Automatic recognition of low-level and high-level surgical tasks in the Operating Room from video images

Florent Lalys

N° d'ordre : 162 **ANNÉE 2012**

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Génie Biologique et Médical

Ecole doctorale : Vie-Agro-Santé

présentée par

Florent LALYS

préparée à l'unité de recherche U746 VisAGeS - Irisa
Vision Action et Gestion d'Information en Santé
Composante universitaire : Science de la vie et de l'environnement

---

Automatic recognition of low-level and high-level surgical tasks in the Operating Room from video images

Thèse soutenue à Rennes
le 03 Mai 2012

devant le jury composé de :

Philippe POIGNET
PU, LIRMM, Montpellier / *rapporteur*

Philippe CINQUIN
PU-PH, TIMC-IMAG, CHU Grenoble / *rapporteur*

Nassir NAVAB
PU, TUM Munich, Allemagne / *examinateur*

Guang-Zhong YANG
PU, Imperial College London, GB / *examinateur*

Xavier MORANDI
PU-PH, CHU Rennes, MediCIS / *examinateur*

Marco WILZBACH
Dr, Carl Zeiss Meditec AG, Allemagne / *examinateur*

Pierre JANNIN
CR INSERM, VisAGeS, MediCIS / *directeur de thèse*

A mes 2 grands-mères…

Remerciements

Lors de cette thèse, j'ai eu l'occasion de rencontrer un certain nombre de personnes que je voudrais remercier ici.

Pour commencer, je tiens à exprimer toute ma gratitude à Pierre Jannin, mon directeur de thèse, qui m'a suivi tout au long de ces 3 ans et qui m'a apporté une aide indispensable, tout en me laissant la liberté nécessaire qu'un chercheur doit avoir. Je tiens à souligner ses qualités humaines ainsi que son implication sans faille dans le monde de la recherche, qui en fait maintenant une personne respectée dans le milieu. Merci de m'avoir poussé à publier et d'avoir accepté les déplacements aux conférences, j'ai pu voyager un peu partout dans le monde et profiter de ces super moments !

Merci à Carl Zeiss Meditec, pour le financement de cette thèse. Travailler en collaboration avec une entreprise privée fut vraiment très appréciable et enrichissant. En particulier, merci à Martin, qui m'a hébergé tout là-bas, au milieu de nulle part à Oberkochen en Allemagne, siège de l'entreprise. Tes conseils scientifiques m'ont beaucoup aidé.

Je remercie le Professeur Philippe Cinquin et le Professeur Philippe Poignet pour avoir accepté d'être rapporteurs de cette thèse. J'adresse également mes remerciements les plus chaleureux aux autres membres du jury pour m'avoir fait l'honneur de juger ce travail et de s'être déplacés de très loin pour certain: le Professeur Xavier Morandi, le Professeur Nassir Navab, le Professeur Guang-Zhong Yang, ainsi que le Docteur Marco Wilzbach.

Je voudrais aussi remercier les membres du service de Neurochirurgie de l'hôpital de Pontchaillou. Principalement Laurent et Claire, que j'ai eu la chance de côtoyer. Deux neurochirugiens humbles, sympathiques, motivés, disponibles et surtout diablement efficaces.

Même si ce sont maintenant des concurrents, des adversaires, que dis-je, des ennemis, je tiens tout de même, non sans une pointe de nostalgie, à remercier mes anciens collègues de l'équipe VisAGeS : Olivier (alias cheville en carton, ou Cortex), Guillaume (alias la gazelle), Pierre M, Aymeric (alias le Don Juan, ou Minus), Sylvain, Camille (désolé pour toutes ces blagues, je conçois qu'être une fille au milieu d'un monde de garçons ne doit pas être facile tous les jours..), Angélique, Elise, Isabelle, Clément (alias mercurochrome). Merci aussi aux anciens membres, qui ont maintenant tous trouvé leur voie : Adrien (alias le schtroumpf pragmatique), qui fut un excellent camarade de sport, Romain (alias el fantôme), Céline, Benoit, Lynda, Yann, Olivier L, Daniel, Vincent, les stagiaires (eux ils n'ont pas de noms…). Aline aussi bien sûr, on pense tous à toi. Merci également à Christian Barillot pour m'avoir accueilli au sein de cette équipe, et d'avoir accepté que je soutienne à l'IRISA. Me pardonneront tous ceux que j'ai oublié de citer ici.

Bien entendu, je n'oublie pas de remercier les membres actuels de la nouvelle équipe Medicis : Tristan (alias Tryphon), Bernard (alias le sage), Maëlle, Tiziano. Euhhhhhhh bah c'est tout….

Une petite pensée également aux profs et responsables de C2i à Rennes 2. Georges notamment. Change pas t'es parfait… ;).

A mes potos. Les vrais de vrais. Ceux qui sont passés du rang de collègues à celui d'amis (ça c'est joliment dit…). Brivaël (alias Nikola) bien sûr, mon ancien co-bureau. Merci d'abord pour ton soutien et ton sens du service, dont beaucoup devrait s'inspirer. Merci aussi pour les pauses cafés à la cafet de la fac de médecine, les pauses cafés à la bibliothèque de la cafèt de médecine, les pauses cafés à l'IRISA, les pauses cafés à Rennes 2, les autres discussions autour d'un café, le café. Alex (alias yocto, ou bubuntu boy), un ingénieur, un pur et dur, mais avant tout un mec généreux. Allé, je peux le dire pour toi aujourd'hui : vive le vélo, Ubuntu, La Réunion, et le PSG (enfin le QSG..). Germain (alias Makumba 67, ou Jems pour les intimes) pour avoir donné un gros coup de boost à l'équipe lors de ton passage. Dans notre milieu, des mecs bons scientifiquement avec qui tu te fends la gueule, ça court pas les rues. Avec le vieux abadie, on formait une bonne équipe hein ? Omar (alias le playboy), autre ancien co-bureau, qui s'est évaporé du jour au lendemain, mais qui je suis sûr va ressortir un jour ou l'autre de sa tanière. Traine pas quand même, le temps passe. Ammar (alias le terroriste), et son humour si spécial mais si appréciable. Même si tu t'es exilé, on ne t'oublie pas par ici ! Bacem (alias Kamel Wali, baçou, ou le boiteux), parce que s'il y a bien un mec qui mérite d'être Docteur dans cette équipe, c'est toi. Le Davido (alias Sergio Busquets, ou le rital), mon nouveau co-bureau. Un mec du tiéquar quand même, ça se respecte.

Une petite pensée également aux matchs/défoulements sur les terrains de foot de beaulieu, de Cap Malo, sur le terrain de squash des Gayeulles, à la la table de ping-pong de l'Irisa, aux tables de billard de la cafet, ou dans les bassins des piscines de Rennes. Mine de rien, tout ça contribue activement à l'épanouissement professionnel !

Il y a bien évidemment une vie au-delà du travail. Je remercie donc ma famille : mes vieux (euh mes parents, pardon), ma sist', mes grands-pères.

Je souhaite enfin et surtout remercier Fanny. Tu es une fille en or. La personne qui m'a le plus appris humainement ces dernières années. Merci pour tout, et pour le reste ! On a vécu de magnifiques choses ensembles, et c'est loin d'être terminé…

Table of contents

List of Figures

List of tables

Lexical

Acronym	Signification
ADL	Analyse Discriminante Linéaire
BVW	Bag-of-Visual-Word
CAO	Chirurgie Assistée par Ordinateur
CBIR	Content-based Image Retrieval
CRF	Conditional Random Field
DBN	Dynamic Bayesian Network
DCT	Discrete Cosine Transform
DoG	Difference of Gaussian
DTW	Dynamic Time Warping
FESS	Functional Endoscopic Sinus Surgery
FPS	Frame Per Second
FRR	Frequency Recognition Rate
GLOH	Gradient Location-Orientation Histogram
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradient
HSV/HSL	Hue Saturation Value/Lightness
IOL	Intra-operative Ocular Lens
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
LESH	Local Energy based Shape Histogram
MEMM	Maximum Entropy Markov Model
MI	Mutual Information
MIS	Minimally Invasive Surgery
OR	Operating Room
ORL	Otorhinolaryngology
PCA	Principal Component Analysis
PPV	Plus Proche Voisin
RFE	Recursive Feature Elimination
RFID	Radio-Frequency IDentification
RGB	Red Green Blue
RN	Réseaux Neurones
ROI	Region Of Interest
SPM	Surgical Process Model
SSD	Sum of Square Difference
SVD	Singular Value Decomposition
SVM	Support Vector Machine

PART I

INTRODUCTION AND RELATED WORK

In Chapter I, I introduce the context of this thesis with a particular focus on Computer-Assisted Surgery. Then, Chapter II is a methodological review of the literature on the creation and analysis of Surgical Process Models, around which this thesis is organized. Both Chapters will permit to introduce the motivations and the problematic of this research.

Chapter I. Introduction

I.1. Présentation de l'équipe VisAGeS

Cette thèse s'est déroulée au sein de l'équipe VisAGeS (VISion, Action, et Gestion d'informations En Santé), rattachée à l'IRISA (UMR CNRS 6074), et commune à l'INRIA (Institut de Recherche en Informatique et Automatique, http://www.inria.fr), l'université de Rennes I (http://www.univ-rennes1.fr) et l'INSERM (Institut National de la Santé et de la Recherche Médicale, http://www.inserm.fr) puis au sein de l'équipe MediCIS (Modélisation des connaissances et procédures chirurgicales et interventionnelles, http://www.medicis.univ-rennes1.fr), équipe INSERM au sein de l'UMR 1099 Laboratoire du Traitement de Signal et de l'Image, Université de Rennes 1.

Les activités de l'équipe VisAGeS (https://www.irisa.fr/visages) concernent le développement de nouveaux algorithmes dédiés à l'analyse d'images médicales et à leurs intégrations dans la salle d'opération. Elles se situent aussi dans le développement de nouveaux algorithmes de traitement de l'information et des interventions assistées par ordinateur dans le contexte des pathologies du système nerveux central. Les travaux de l'équipe sont plus particulièrement centrés sur la conception de la salle d'opération du future, une meilleure compréhension des pathologies du cerveau à différentes échelles. Trois principales thématiques liées à des domaines d'application différents se dégagent des travaux de l'équipe.

La première thématique, portée par Christian Barillot, s'intéresse aux biomarqueurs d'imagerie dans les pathologies du cerveau. Plus particulièrement, des workflows de traitement d'images et d'analyse sont mis en œuvre pour extraire et exploiter des biomarqueurs d'imagerie. Les champs de recherche sont variés, de la physique médicale à l'acquisition des données, en passant par le traitement, l'analyse et la fusion des images médicales. Avec ces outils, les applications médicales concernent la sclérose en plaque, la maladie de Parkinson, la neuro-pédiatrie, l'Arterial Spin Labeling et la morphométrie 3D endocrânienne.

La seconde thématique, portée par Bernard Gibaud, se situe dans la gestion d'informations en neuro-imagerie. L'idée de ces travaux est d'annoter des données image ainsi que les méta-informations en découlant en se référant à des ontologies de domaine, dans le but de rendre explicite leur sémantique. Ces travaux facilitent le partage et la réutilisation des données pour des recherches en neuro-imagerie.

La troisième et dernière thématique, portée par Pierre Jannin, se porte sur la neurochirurgie assistée par des modèles. Devant l'apparition de nombreux outils dans les salles d'opération, des nouveaux systèmes assistés par ordinateur sont crées dans le but d'aider le chirurgien dans la tâche opératoire. Ces systèmes peuvent être basés sur des informations préopératoires et intra-opératoires ainsi que sur des modèles de procédures décrivant le scénario chirurgical. Dans ce contexte, les objectifs de ces travaux sont d'aider le planning préopératoire, par exemple en Stimulation Cérébrale Profonde, d'étudier les déformations intra-opératoires dues au brain-shift et de créer des modèles de procédures basés sur le processus chirurgical ou sur des analyses cognitives des chirurgiens.

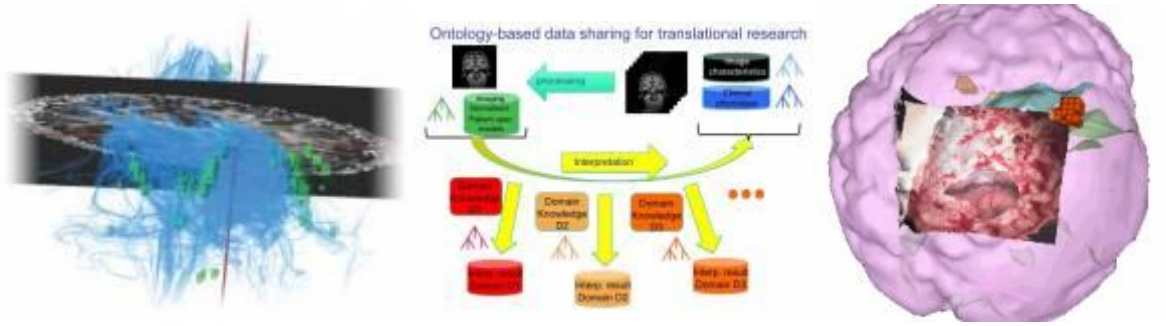


Figure 1 - Trois thématiques de l'équipe VisAGeS

Les activités de l'équipe MediCIS (Modélisation des connaissances et procédures chirurgicales et interventionnelles, http://www.medicis.univ-rennes1.fr) regroupent les activités de Bernard Gibaud et Pierre Jannin pour la conception de systèmes d'aide à la décision chirurgicale par l'étude et la construction de modèles de connaissances et de procédures. Ces modèles sont étudiés par des approches à la fois symboliques et numériques.

I.2. Chirurgie assistée par ordinateur

I.2.a. Contexte

A l'heure actuelle, la présence des nouvelles technologies dans le domaine médical se fait ressentir. La salle d'opération, cœur de la prise en charge des patients à l'hôpital, a subie de profondes transformations pour évoluer vers un environnement complexe et riche en technologie de pointe. Les technologies de l'informatique sont maintenant essentielles à son bon fonctionnement. Celles-ci sont de plus en plus utilisées au cours de l'intervention chirurgicale : du planning pré-opératoire à l'évaluation post-opératoire, en passant bien sûr par l'aide intra-opératoire. C'est dans ce contexte que sont nés les systèmes de **Chirurgie Assistée par Ordinateur (CAO)**. La CAO est définie comme l'ensemble des systèmes aidant le praticien dans la réalisation de ses gestes diagnostiques et thérapeutiques.

En phase pré-opératoire, ces systèmes fournissent un accès aux images multimodales et aux informations des patients. Ils permettent ainsi de préparer, voire de simuler, un scénario chirurgical propre à chaque patient. Pendant la chirurgie, ils apportent une interface de visualisation en intégrant ces différentes données. Des robots peuvent aussi assister, à différents degrés (aide passive, semi-active ou active) le geste chirurgical selon le degré d'indépendance du robot vis-à-vis de la tâche chirurgicale. En phase post-opératoire, ils fournissent des outils d'aide pour l'analyse et l'évaluation de la procédure.

Ces systèmes de CAO ont donc pour avantage d'aider à la prise de décision et d'améliorer la prise en charge du patient. La pertinence clinique de ces nouveaux outils technologiques étant partiellement démontrée, les enjeux actuels résident donc dans la création d'outils pour une prise en charge chirurgicale codifiée, sécurisée, et optimisée à chaque patient. Ces questions ont été discutées par Cleary et al. (2005), Rattner et Park (2003), Xiao et al. (2008), Satava et al. (2001), Avis (2000) ou Gorman et al. (2000).

Pour une optimisation de ces systèmes, deux aspects sont fondamentaux dans la CAO : l'établissement de **modèles spécifiques au patient** et la **modélisation des procédures**.

I.2.b. Modèles spécifiques aux patients

Le chirurgien a besoin d'un ensemble d'images pré-opératoires multimodales pour tenir compte de la complexité anatomique, physiologique et métabolique des cibles et de l'environnement chirurgical. Ces modèles établissent un lien direct entre le patient en salle d'opération, dans le référentiel de la salle, et ses multiples images.

Durant la phase pré-opératoire de planning, le chirurgien a besoin d'établir une cartographie spécifique de son patient. Celle-ci est établie à partir des images anatomiques (Scanner CT, IRM) ou fonctionnelles (IRM de diffusion, TEP, TEMP, IRMf) spécifiques au patient. Des outils de traitement d'images sont couramment appliqués pour extraire les informations pertinentes pour le chirurgien (

Figure 2), comme la segmentation d'une tumeur ou la visualisation des faisceaux de fibres. Lorsque plusieurs séquences d'images sont acquises, un recalage, linéaire et/ou non-linéaire, est nécessaire pour les regrouper dans un repère commun et ainsi permettre la cartographie. De même, des données propres au patient (âge, sexe, pathologies, etc...) peuvent être intégrées dans ces modèles pour aider à la prise de décision ou créer des groupes homogènes de patient.

Pendant la chirurgie, l'opérateur doit maitriser la relation spatiale entre le patient et son modèle. Cette mise en relation des deux repères peut être effectuée par un repérage anatomique de points ou de surfaces dans l'espace du patient reportés ensuite dans le modèle. Premier exemple de repérage 3D en neurochirurgie : la chirurgie stéréotaxique, un cadre fixé à la tête du patient définit un repère commun. Deuxième exemple, incontournable : la neuronavigation. Cela est rendu possible grâce à des localisateurs installés en périphérie de la table d'opération, autour du patient et agissant comme des systèmes GPS (Global Positioning System). Ceux-ci permettent de localiser, en temps réel, des cibles positionnées sur des objets physiques et de connaître leur position dans le repère du modèle du patient. La chirurgie est alors guidée par l'imagerie. Grâce à ce type de système, la chirurgie devient plus sûre, le chirurgien dispose d'une aide considérable pour se repérer et éviter ainsi de potentielles erreurs. Troisième et dernier exemple : la réalité augmentée. Celle-ci s'attache à surajouter à l'environnement de la chirurgie (réel) des informations numériques préalablement acquises (virtuelles). Couplé aux modèles spécifiques au patient, la réalité augmentée fait partie des nouveaux systèmes qui apportent une aide non-négligeable aux chirurgiens.

Au-delà de cette approche de modèles spécifiques au patient, l'optimisation des nouveaux systèmes de CAO passe par la mise en place de la modélisation des procédures chirurgicales.

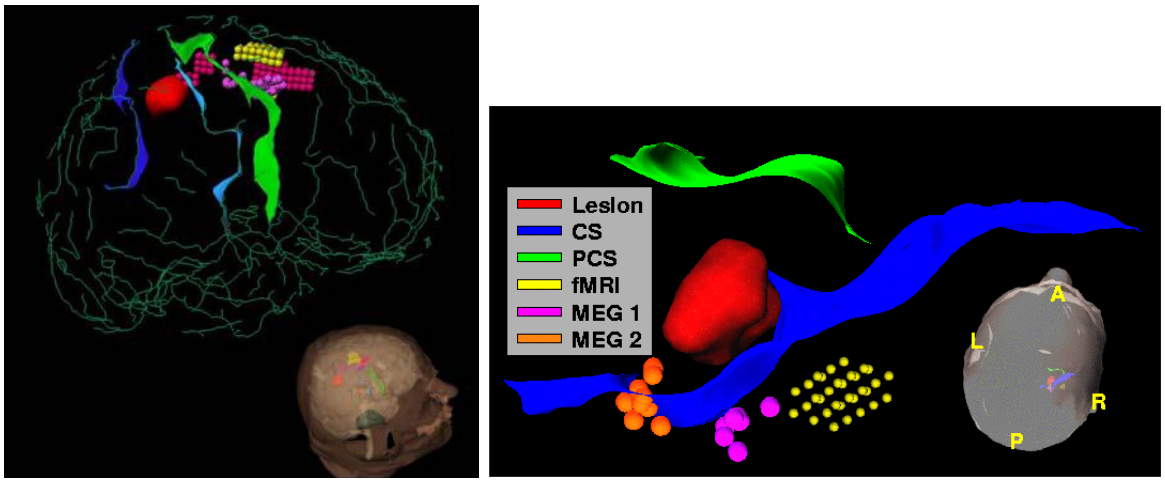


Figure 2 - Exemple de modèles spécifiques au patient en neurochirurgie tumorale.

I.2.c. Modèles de procédures chirurgicales

Il est important de concevoir une salle d'opération qui offre au chirurgien et à son équipe une facilité de travail par un accès aux images, informations ou outils disponibles. Ainsi une connaissance du flux d'actions est primordiale pour spécifier et définir la salle d'opération du futur (Cleary et al. 2005). La procédure est décrite de manière principalement symbolique comme une succession d'étapes et d'actions réalisées avec différents outils et selon différentes techniques. La connaissance du flux d'action à travers des modèles de procédures a donc pour but de créer une nouvelle génération de systèmes de CAO qui s'appuie sur une formalisation du processus chirurgical et des connaissances. Cette formalisation peut être construite à partir soit des procédures réalisées par des chirurgiens, soit d'un consensus d'experts. Elle doit permettre de décrire le plus précisément possible les procédures en respectant le déroulé de la chirurgie et en se rapprochant de la réalité.

Les objectifs de ces modèles sont multiples. Ceux-ci doivent permettre d'expliquer pourquoi la procédure suit ce déroulé, c'est-à-dire savoir pourquoi à un moment donné de la procédure, une activité particulière est réalisée. Ils doivent également aider à distinguer les différences entre des procédures. Ces modèles doivent aussi aider à prévoir quelle sera l'étape suivante lors d'une procédure, et de manière plus globale quelle sera la procédure utilisée pour un patient donné. Lors de la phase de planning, le chirurgien pourra se référer soit à des scénarios-types issus de combinaisons et fusions de cas, soit à des cas semblables déjà effectués.

La modélisation des procédures chirurgicales a été introduite pour des applications multiples. Un exemple est la visualisation sur écran des informations pertinentes pour le chirurgien. Celles-ci peuvent être adaptées et triées tout au long de l'acte chirurgical en fonction du modèle de la procédure. Un deuxième exemple est celui du développement d'outils d'apprentissage de la chirurgie. En effet, la formalisation permet de définir, entre autre, une terminologie adaptée pouvant être réutilisée et servir de base pour des descriptions explicites de la procédure. Cela pourrait contribuer aux progrès des systèmes assistés par ordinateur dans la salle d'opération (Lemke and Vannier, 2006; Cleary et al, 2005; Burgert et al, 2006a, 2006b). D'autres méthodes venant de domaines non médicaux ont été adaptées à l'environnement chirurgical. Dickhaus et al. (2004) ont démontré que la méthode BPR (Business Process Reengineering) pouvait aider les systèmes assistés par ordinateur. Lemke and Berliner (2007) ont introduit un concept pour l'interopérabilité des données entre composants des systèmes chirurgicaux. Ce type de système fut conçu pour améliorer les communications et le management des images dans la salle d'opération.

La modélisation des procédures chirurgicales a aussi motivé le développement de suppléments dans le format d'images médicales DICOM (Digital Imaging and COmmunications in Medicine) (Lemke, 2007). DICOM définit la représentation, le transfert, le stockage et la génération des données images. Ainsi, Burgert et al. (2007) ont proposé une analyse basée sur les workflows chirurgicaux qui aident à la prise en charge des informations du patient en plus des données images. Des modèles géométriques furent utilisées, représentant les différents aspects du workflow chirurgical, comme les structures anatomiques, les outils chirurgicaux, etc. Cette étude a expliqué le processus de spécification dans le but de fournir un template pour la définition de nouvelles classes DICOM. Les workflows ont enfin été introduits pour assister les systèmes de réalité augmentée (Navab et al., 2007) et pour les nouveaux challenges en télé-médecine (Kaufman et al., 2009).

Une attention particulière a récemment été donnée à la création de modèles de procédures chirurgicales. La modélisation du processus chirurgical est ainsi la base des nouveaux systèmes de CAO autour duquel s'inscrit cette thèse. Le chapitre suivant va permettre d'effectuer un état de l'art complet sur les modèles de processus chirurgicaux, i.e. *Surgical Process Model* (SPM), et d'introduire en détail la problématique de cette thèse.

Réferences

- ✓ Avis NJ. Virtual environment technologies Minim Invasive Ther Allied Technol. 2000; 9(5): 333-40.
- ✓ Burgert O, Neumuth T, Lempp F, Mudunuri R, Meixensberger J, Strauß G, Dietz A, Jannin P, Lemke HU. Linking top-level ontologies and surgical workflows. Int J Comput Assisted Radiol Surg. 2006a; 1(1): 437-8.
- ✓ Burgert O, Neumuth T, Fischer M, Falk V, Strauss G, Trantakis C, Jacobs S, Dietz A, Meixensberger J, Mohr FW, Korb W, Lemke HU. Surgical workflow modeling. MMVR. 2006b.
- ✓ Cleary K, Chung HY, Mun SK. OR 2020: The operating room of the future. Laparoendoscopic and Advanced Surgical Techniques. 2005; 15(5): 495-500.
- ✓ Dickhaus CF, Burghart C, Tempany C, D'Amico A, Haker S, Kikinis R, Woern H. Workflow Modeling and Analysis of Computer Guided Prostate Brachytherapy under MR Imaging Control. Studies Health Technol Inform. 2004; 98: 72-6.
- ✓ Gorman PJ, Meier AH, Rawn C, Krummel TM. The Future of Medical Education Is No longer Blood and Guts, It Is Bits and Bytes. Am J Surg. 2000; 180: 353-5.
- ✓ Jannin P. De la neurochirurgie guidée par l'image, au processus neurochirurgical assisté par la connaissance et l'information. HDR de l'université de Rennes I, Faculté de Médecine. 2005.
- ✓ Kaufman DR, Pevzner J, Rodriguez M, Cimino JJ, Ebner S, Fields L, Moreno V, McGuiness C, Weinstock RS, Shea S, Starren J. Understanding workflow in telehealth video visits: Observations from the IDEATel project. J Biomed Informatics. 2009; 42(4): 581-92.
- ✓ Lemke HU and Berliner L. Specification and design of a therapy imaging and model management system (TIMMS). SPIE medical imaging - PACS and Imaging Informatics. 2007; 6516:651602.
- ✓ Lemke HU and Vannier MW. The operating room and the need for an IT infrastructure and standards. Int J Comput Assisted Radiol Surg. 2006; 1(3): 117-22.
- ✓ Lemke HU. Summary of the White Paper of DICOM WG24: DICOM in Surgery. SPIE Medical Imaging 2007 – PACS and Imaging Informatics. 2007: 6516:651603.
- ✓ Mueller ML, Ganslandt T, Frankewitsch T, Krieglstein CF, Senninger N, Prokosch HU. Workflow analysis and evidence-based medicine: towards integration of knowledge-based functions in hospital information systems. Proc AMIA Symp. 1999; 330-4.
- ✓ Navab N, Traub J, Sielhorst T, Feuerstein M, Bichlmeier C. Action-and workflow-driven augmented reality for computer-aided medical procedures. Computer Graphics. 2007; 27(5): 10-4.
- ✓ Qi J, Jiang Z, Zhang G, Miao R, Su Q. A surgical management information system driven by workflow. IEEE conf service operations and logistics, and informatics. 2006; 1014-8.
- ✓ Rattner WD, Park A. Advanced devices for the operating room of the future. Seminars in laparoscopic surgery. 2003; 10(2): 85-9.
- ✓ Satava RM. Accomplishments and challenges of surgical simulation dawning of the next generation surgical education. Surg Endosc. 2001; 15: 232-41.
- ✓ Trevisan DG, Vanderdonckt J, Macq B, Raftopoulos C. Modeling interaction for image-guided procedures. SPIE medical imaging Visualisation, Image-guided procedures and display. 2003: 5029 ; 108
- ✓ Xiao Y, Hu P, Moss J, de Winter JCF, Venekamp D, MacKenzie CF, Seagull FJ, Perkins S. Opportunities and challenges in improving surgical work flow. Cognition Technology. 2008; 10(4): 313-21.

Chapter II. Review on creation and analysis of Surgical Process Models

II.1. Introduction

II.1.a. Context

As introduced in the previous Chapter, the Operating Room (OR) has particularly undergone significant transformations to evolve into a highly complex and technologically rich environment. Computer technologies are now essential and increasingly used throughout the intervention, from pre-operative planning to post-operative assessment. Computer-Assisted Surgery (CAS) (or Computer-assisted Intervention-CAI) systems have now a vital role in current surgeries performance. Following the progress of models of surgical procedures, the necessity is now to understand the process of the surgery in order to better manage the new generation of CAS systems. A new terminology has been defined around this aspect of models of surgical procedures. The term *surgical workflow* has been defined by Jannin and Morandi (2007). It follows the glossary of the Workflow Management Coalition (WFMC 1999), defining a surgical workflow as "the automation of a business process in the surgical management of patients, in whole or part, during which documents, information, images or tasks are passed from one participant to another for action, according to a set of procedural rules". This idea of decomposing the surgery into a sequence of tasks was first introduced by MacKenzie et al. (2001), and was later formalized by Neumuth et al. (2007). They defined a *Surgical Process* (SP) as a set of one or more linked procedures or activities that collectively realize a surgical objective within the context of an organizational structure defining functional roles and relationships. This term is generally used for denominating a surgical procedure course. They also defined a *Surgical Process Model* (SPM) as a simplified pattern of a SP that reflects a predefined subset of interest of the SP in a formal or semi-formal representation. It is related to the performance of a SP with support of a workflow management system. SPMs have been first introduced for supporting the surgical intervention thanks to a model of the surgery progress. Indeed, the precondition of a computer supported surgical intervention is the specification of the course model describing the operation to be performed (Cleary et al., 2005). Typically, even if every surgery is different, the same type of procedure shares common sequences of states that can be extracted. Being able to extract information such as activities, steps or adverse events in a surgery and having the possibility to rely on a surgery model is therefore a powerful tool to help surgeons.

SPM methodology could be crucial for future components of CAS systems since it may have a direct impact on many aspects of the procedure. The use of SPM may prove its efficiency for facilitating the surgical decision-making process as well as improving the pre-operative human-computer interface and medical safety.It would have direct impact on the process of care-based

decisions. It could find its applications in the anticipation of patient positioning, the optimisation of operating time, the evaluation of surgeons, tools, or the analysis of technical requirements. We propose in this Chapter a first methodological review of the literature focusing on the creation and the analysis of SPMs.

II.1.b. Search methodology

The review was done according to a search on Google Scholar on the specific keywords: "surgical process model", "surgical process analysis", "surgical ontology", and "surgical workflow analysis". In addition to the Google Scholar results, we added another list of possible citations that were extracted from the references of the publications. We included articles published in peer-reviewed journals as well as full papers published in international conference proceedings that were concerned with the use of SPM. International conferences proceedings were included because the area is very recent resulting in many conference publications but few peer-reviewed journals. Only English language has been accepted. Included researches have been published from 1998 until December 2011. In order to get an overview of publications that focused on the creation and analysis of SPMs, we were interested in studies that model the procedural approach, i.e. works that took into account the sequential aspect of the surgical procedure. Moreover, we were interested in works that focused at least one part of their analysis on the act of surgery, beginning when the surgeon performs the first task on the patient and ending when the surgeon closes with the suture. When a project has been published multiple times with no change in the dedicated elements of the diagram, either the more recent or the one in the best journal was kept. The entire process of selection is shown on **Figure 3**. From a first selection of N=250 publications, a total of N=43 publications were finally conserved for full-text review.

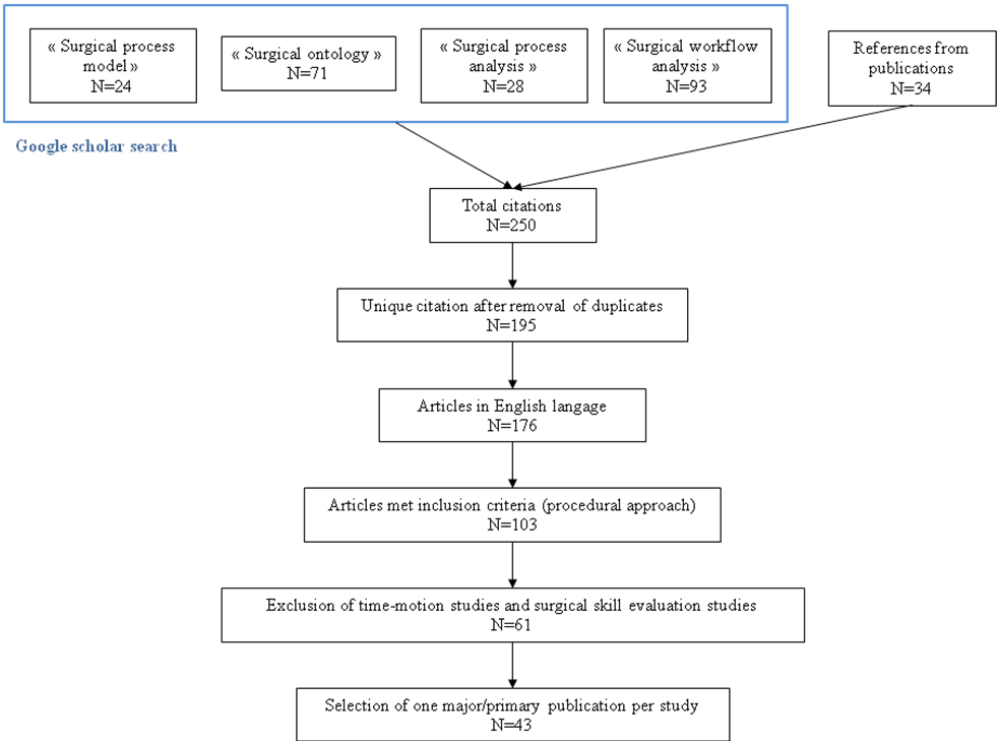


Figure 3 - Process used in the selection of publications for full-text review.

Figure 4 shows the results of the Google scholar results only before the process of selection. We can see that the area of creation and analysis of SPMs is very recent. It has particularly evolved from 2007 that shows the recent evolution of the domain.

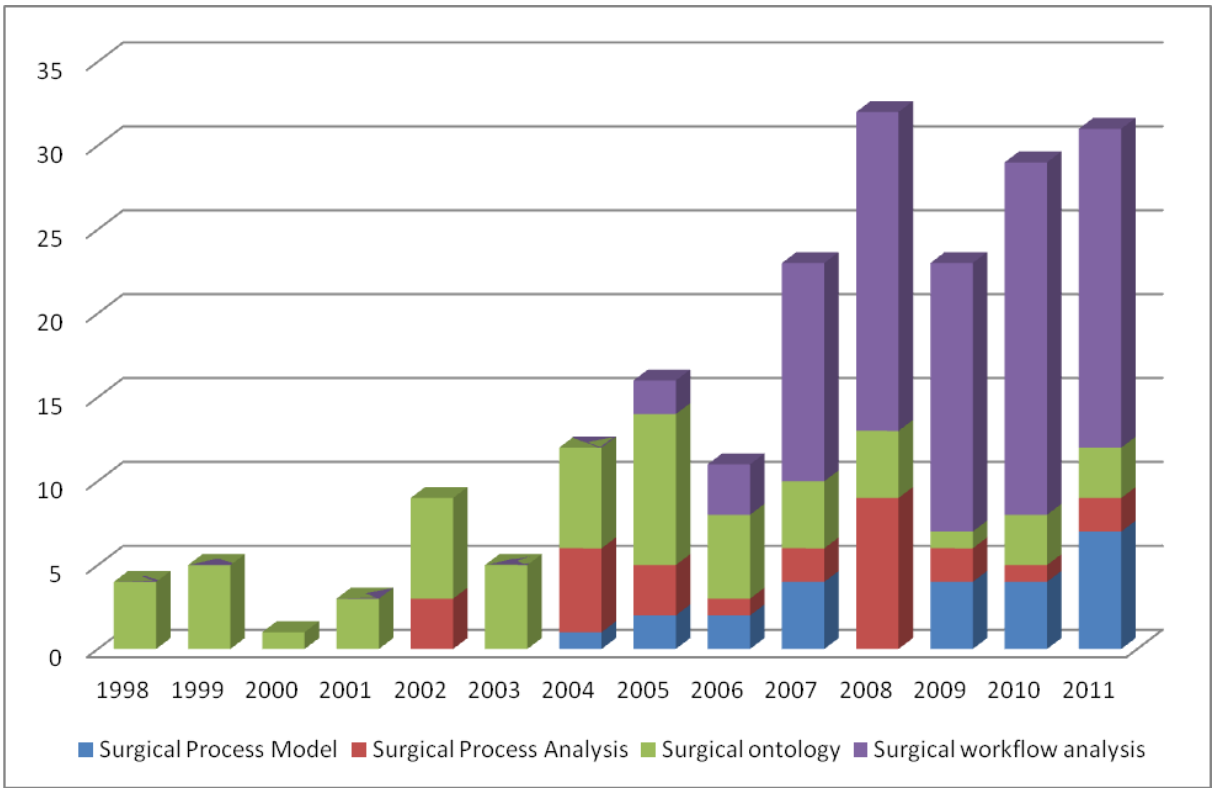


Figure 4 - Evolution of the number of papers in the field from 1998 to December 2011

II.2. SPM methodology

In order to clarify the review and the discussions, we propose a model for describing and classifying the methods using five components and their corresponding elements (**Figure 5**). Each of the five components addresses one major aspect of the SPM methodology, and every element that is resulting can be instantiated with its set of possible values. The first component is the modelling, where the goal is to describe the work-domain of the study and its formalism. The next component is the acquisition which is the second step of a SPM methodology that allows the acquisition of data by human observations or by sensor systems. The third one is the analysis that tries to make the link between data acquisition and the information that we want to model. Another component specifies the different applications of the systems based on SPMs and finally the last component describes the different kind of validation and evaluation that are conducted for assessing these systems. The whole review is organized according to this diagram. In the following subsections, each component and each element are explained in detail.

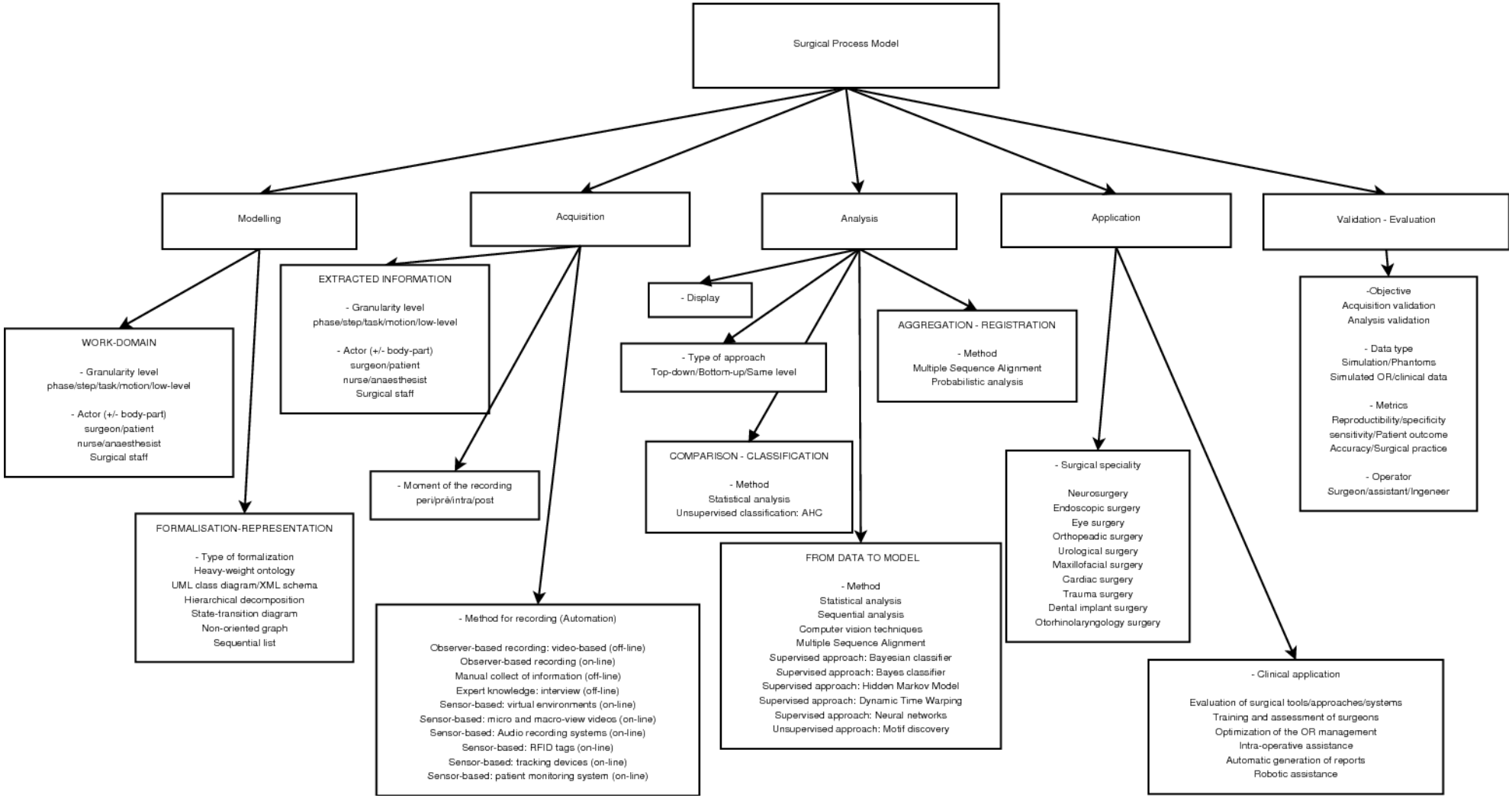


Figure 5 - Overview graph of the field

II.2.a. Granularity level

The whole SPM methodology, and especially the acquisition and modelling component, is organized around the aspect of granularity level. Surgery can be studied at different granularity levels defined as the level of abstraction for describing a surgical procedure. New terms describing the different levels have been introduced and adapted to SPMs for a better standardisation of surgical descriptions. The group of MacKenzie (Cao et al., 1996; Ibbitson et al., 1999; MacKenzie et al., 2001) first proposed a model of the surgical procedure that consists of different levels of granularity: the procedure, the step, the substep, the task, the subtask and the motion. Each (sub)task can be for instance decomposed in various motions and forces primitives. Then they used a hierarchical decomposition for structuring the complex environment and the interaction between the surgical team and new technologies. Because of the large differences of terminology employed by the studied papers, in this Chapter we will use the following terminology for describing the different granularity levels of surgical procedures. The highest level would be the procedure itself, followed by the phases, the steps, the activities, the motions and lastly all other low-level information such as position of instruments or images (**Figure 6**). One assumption is that each granularity level describes the surgical procedure as a sequential list of events, except for the surgical procedure itself and for lower-levels where information may be continuous. The **motion** is defined as a surgical task involving only one trajectory but with no semantics. This granularity level would be identical to the definition of "dexemes" by Reiley and Hager (2009). The **activity** is defined as a surgical task with a semantic meaning involving only one surgical tool, one anatomical structure and one action, as formalized by Neumuth et al. (2006). This level would be identical to the "surgemes" definition. At a higher level, a **step** is defined as a sequence of activities toward a surgical objective, which have been often called "task" in the literature. Finally, the **phase** level is defined as a sequence of tasks at a higher level that may involve other members of the surgical staff. It would be identical to the surgical episode of Lo et al. (2003).

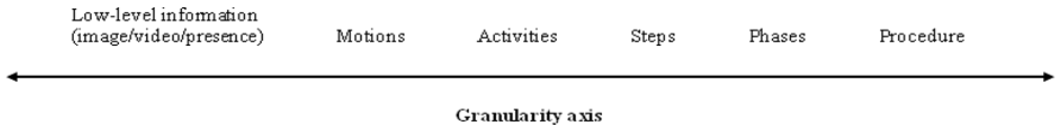


Figure 6 - Different levels of granularities of a surgical procedure.

II.2.b. Modelling

The first component describes and explains the work-domain of the study, i.e. what is studied and what is modelled. Two information are needed: 1) the granularity level of the surgical information and 2) the operator. A third element completes this component: 3) the formalization of the information. In many cases, a phase of formalization is necessary for representing the collected knowledge before the analysis process. Knowledge acquisition is the process of extracting, structuring and organizing knowledge from human experts. It has to be part of an underlying methodology and incorporate a strong semantic aspect.

Granularity level

Similar to the data acquisition process, information that is studied (i.e. information that is modelled) is disposed on the granularity axis previously defined. The activities have been mainly investigated, but all granularity levels have been studied. At the highest level, the global procedure has been studied (Bhatia et al., 2007; Hu et al., 2006; Sandberg et al., 2005; Xiao et al., 2005), as well as the phases (Ahmadi et al., 2007; James et al., 2007; Katic et al., 2010; Klank et al., 2008; Lo et al., 2003; Nara et al., 2011; Padoy et al., 2007, 2008, 2010; Qi et al., 2006; Suzuki et al., 2010), the steps (Blum et al., 2008; Bouarfa et al., 2010; Fischer et al., 2005; Jannin et al., 2003, 2007; Ko et al., 2007; Lemke et al., 2004; Malarme et al., 2010; and the motions (Ahmadi et al., 2009; Lin et al., 2006; Nomm et al., 2008). Some studies integrated two or more of these granularity levels in their modelling (Burgert et al., 2006; Ibbotson et al., 1999; MacKenzie et al., 2001; Münchenberg et al., 2001; Xiao et al., 2005; Yoshimitsu et al. 2010). No low-level information was considered here.

Operator

Information that is studied involves one or many of the actors of the surgery: the operator can be the surgeon, the nurses, the anaesthetist, the patient or many of these operators

Formalization

Formalization is necessary for allowing automated treatment and processing by computers. It is also necessary for bottom-up approaches to have a representation of the sequence of the surgery trough ontologies or simple list of phases/steps/activities. At the highest level, we find the heavy-weighted ontologies, which have been used to represent the detailed context of a SPM study. Then, in the category of light-weighted ontologies, we find UML class diagrams along with XML schema. Both approaches define entities and relation between these entities. We then find all 2D graphs representations, with the hierarchical decompositions, the state-transition diagram and the non-oriented graphs. Lastly, at the lower level, simple sequential list were also used, proposing an ordered list of word for representing one or many levels of granularity of the surgery (**Figure 7**).

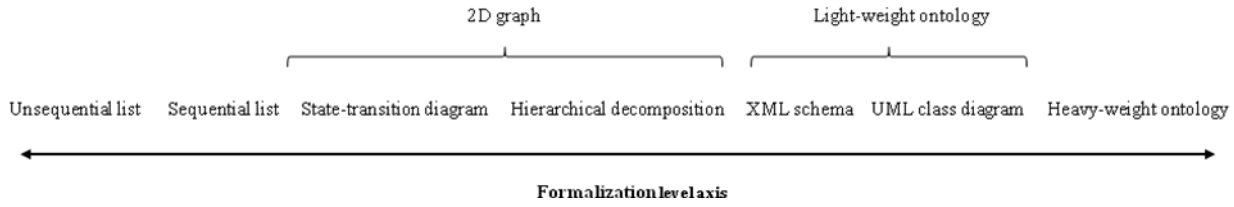


Figure 7 - Different levels of formalisation of the surgery.

II.2.c. Data Acquisition

The second component of the diagram is the acquisition, i.e. the collection of data on which the models are build. Four main elements can be defined for the acquisition process: 1) the level of granularity of the surgical information that is extracted, 2) the operator(s) on which information are extracted, 3) the moment when the acquisition is performed, and 4) the recording method. This section is divided according to these four elements.

Granularity level

The level of granularity of the surgical information that is extracted allows characterizing the acquisition, as it determines in which detail the SP is recorded. Studies have focused on the recording of the entire procedure (Sandberg et al., 2005), the phases (Qi et al., 2006), the steps (Burgert et al., 2006; Fischer et al., 2005; Lemke et al., 2004), the activities (Forestier et al., 2011; Meng et al., 2004; Neumuth et al., 2006, 2009, 2001a, 2011b; Riffaud et al., 2011) and the motions (Kragic and Hager, 2003). But efforts have been particularly made on the extraction of low-level information from the OR: images (Jannin et al., 2003, 2007; Münchenberg et al., 2001), videos (Bhatia et al., 2007; Blum et al., 2008; Klank et al., 2008; Lo et al., 2003; Speidel et al., 2008), audio, position data (Houliston et al., 2011; Katic et al., 2010; Ko et al., 2007; Sudra et al., 2007), trajectories (Ahmadi et al., 2009; Ibbotson et al., 1999; Lin et al., 2006; Miyawaki et al., 2005; Nara et al., 2011; Nomm et al., 2008; Yoshimitsu et al., 2010), information of presence/absence of surgical tools (Ahmadi et al., 2007; Bouarfa et al., 2010; Padoy et al., 2007) or vital signs (Xiao et al., 2005). Several of these low-level information can also be combined (Agarwal et al., 2007; Hu et al., 2006; James et al., 2007; Malarme et al., 2010; Padoy et al., 2008, 2010; Suzuki et al., 2010).

Operator

Surgery always directly involves several operators. All staff members can have an impact on the surgery and their roles and actions can be studied. The most important operator is of course the surgeon, which is performing the surgery or surgical tools when positions, trajectories or information of presence of surgical tools are extracted. But other operators can be involved: the nurse (Miyawaki et al., 2005; Yoshimitsu et al., 2010) for trajectories data extraction, the patient (Agarwal et al., 2007; Hu et al., 2006; Jannin et al., 2003, 2007; Münchenberg et al., 2011; Sandberg et al., 2005; Suzuki et al., 2010; Xiao et al., 2005) for images or vital signs extraction, or the anaesthetist (Houliston et al., 2011). Global studies on the entire surgical staff have also been proposed (Agarwal et al., 2007; Bhatia et al., 2007; Fischer et al., 2005; Hu et al., 2006; Lemke et al., 2004; Nara et al., 2011; Qi et al., 2006; Sandberg et al., 2005; Suzuki et al., 2010), where the surgeon, the nurses and possibly the anaesthetist are concerned. For tracking systems, we can also specify, when it is clearly defined, the corresponding human body parts involved, such as hand, eye, forehead, wrist, elbow, and shoulder.

Moment of acquisition

The moment when the data acquisition is performed (timeline) is also vital information for discriminating acquisition techniques. The acquisition most of the time extracts data from intra-operative recordings, but for it can also be post-operative acquisitions (*retrospective*) in the case of observer-based recording from video or some tracking systems, or pre-operative acquisitions (*prospective*) in the case of manual collect of information. Additionally, the term peri-operative generally refers to the three phases of the surgery. Some acquisitions integrate all of these three phases for having information from the entire patient hospitalization process.

Methods for recording

Two main recording approaches have been proposed: observer-based and sensor-based approaches. Observer-based approaches are performed by one person who needs a certain surgical background. For off-line recording, the observer used one or multiple videos from the OR to retrospectively record the

surgical procedure (Ahmadi et al., 2007, 2009; Bouarfa et al., 2010; Fischer et al., 2005; Ibbotson et al., 1999; Lemke et al., 2004; MacKenzie et al., 2001; Malarme et al., 2010; Padoy et al., 2007). For on-line recording, the observer is directly in the OR during the intervention (Forestier et al., 2011; Neumuth et al., 2006a, 2006b, 2009, 2011; Rifaud et al., 2011). Lemke et al. (2004) first presented interests of studying OR using on-line observer-based approaches to progress in both ergonomic and health economic.

Sensor-based approaches have been developed for automating the data acquisition process and/or for finer granularity descriptions. The principle is to extract information from the OR thanks to one or multiple sensors in an automatic way, and to recognize activities or events based on these signals. Sensors can be of different types, from electrical to optical systems. First, studies have used sensors based on Radio Frequency IDentification (RFID) technologies directly positioned on instruments or on the surgical staff during the intervention to detect the presence/absence of the positions (Agarwal et al., 2007; Houliston et al., 2009). Then, efforts have been made on robot-supported recording (Ko et al., 2007; Kragic and Hager, 2003; Lin et al., 2006; Münchenberg et al., 2001), including surgeon's movements and instruments use. Robots have been used as a tool for automatic low-level information recordings. Tracking systems (Ahmadi et al., 2009; James et al., 2007; Katic et al., 2010; Miyawaki et al., 2005; Nara et al., 2011; Nomm et al., 2008; Sudra et al., 2008; Yoshimitsu et al., 2010) have also been used in various studies, with eye-gaze tracking systems positioned on surgeons or staff members tracking devices. Other types of methods have also been tested for recording information: Patient monitoring systems (Agarwal et al., 2007; Hu et al., 2006; Sandberg et al., 2005; Xiao et al., 2005), or audio recording systems (Agarwal et al., 2007; Suzuki et al., 2010). Lastly, the use of on-line video-based recording, sometimes combined with other data acquisition techniques, has particularly received increased attention recently (Bhatia et al., 2007; Blum et al., 2008; Hu et al. 2006; James et al., 2007; Klank et al., 2008; Lo et al., 2003; Padoy et al., 2008, 2010; Speidel et al., 2008; Suzuki et al., 2010), with either macro-view videos recording the entire OR or micro-view videos such as endoscope videos.

Observer-based approaches			**Sensor-based approaches**						
Observer-based recording from video (off-line)	Observer-based recording (on-line)	Manual collect of information	Robot-supported recording (on-line)	Robot-supported recording	On-line video-based recording	Patient monitoring systems	RFID technologies	Tracking systems	Audio recording systems

Table 1 - List of possible data acquisition methods.

II.2.d. Analysis

Analysis methods can be divided into three types: the methods that go from the data to the final model, the methods that aggregate or fuse information and the methods that classify or compare data for extracting a specific parameter. The three approaches are presented in the next subsections. Additionally display methods of the analysis results have been studied to have a visual representation after the analysis process.

From data to model

The challenge here is to use the data collected during the acquisition process to create an individual model (i.e. iSPM) and to make the link between the acquisition process and the modelling. The type of

approach used can be determined by comparing the level of granularity of the acquisition information and of the modelling. Top-down approaches are described as analyses that go from a global overview of the intervention with patient-specific information and a description of high-level tasks (such as phases or steps) to fine-coarse details (such as activities or motions). On the contrary, a bottom-up approach takes as input low-level information from sensor devices and tries to extract semantic high-level information. The methodology employed for either bridging the semantic gap in the case of bottom-up approaches or generalize and formalize individual recordings in the case of top-down approaches is based on statistical, informatics, or data-mining concepts. The level of automation during the creation of the model has to be defined here. The issue is to determine if the model needs a training step or not. This step is needed for assigning classes to the training set. In such cases, the creation of the model isn't fully automatic and may be entirely manual or a mix between human intervention and automatic computation.

Within supervised approaches, simple Bayes classifier and neural networks have been tested in the case of activity/step/phase recognition. For analysing patient vital signs, signal processing tools have been used. In the case of top-down analysis, description logic has been widely studied. Towards more complex models, graphical probabilistic models are often useful to describe the dependencies between observations. Bayesian Networks (BN) have recently proven to be of great interest for such applications, with an extension in the temporal domain using Dynamic BNs (DBN). Temporal modelling allows evaluating the duration of each step and the entire process during the execution. Many time-series models, such as Hidden Markov Model (HMM) (Rabiner, 1989) or Kalman filter models, are particular examples of DBNs. Indeed, HMM, which are statistical models used for modelling non-stationary vector times-series, have been widely used in SPM analysis. Another time-series analysis tool has been often tested with success because of its capacity of temporal registration, the Dynamic Time Warping (DTW) algorithm. Computer vision techniques have also been employed but for extracting high-level information before using supervised approaches. Computer vision techniques allow going from a low-level description of images and videos to high-level semantic meaning. Within unsupervised methods, no extensive works have been done. We only find the motif discovery approach (Ahmadi et al., 2009) that doesn't need any à priori model. Statistical analysis, sequential analysis or trajectories analysis have also been used. Lastly, using text-mining for extracting data and create model has been also tested. The idea is to automatically analyse post-operative procedure reports as well as patient files (Meng et al., 2004).

A SPM whose data acquisition and modelling stay at the same level of granularity is also possible. In such cases, the goal of the analysis is not to create a real model, but to perform either aggregation/fusion or comparison/classification.

Comparison-Classification

The principle is to use SPMs to highlight a specific parameter (i.e. meta-information) that explains differences between populations of patients, surgeons or systems. Simple statistical comparisons (such as average, number of occurrence or standard deviation) have been used (Ibbotson et al., 1999; Riffaud et al., 2010; Sandberg et al., 2005) to compare populations. Similarity metrics have also been proposed by Neumuth et al. (2011a) to be able to compare different SPs. DTW along with K-Nearest Neighbour (KNN) algorithm have been tested within unsupervised approaches (Forestier et al., 2011).

Aggregation-Fusion

The goal here is to create a global model (gSPM) of a specific procedure by merging a set of SPMs. One possibility is to merge similar paths as well as filter infrequent ones to create average SPs for having a global overview of the surgery. Another is to create gSPMs that represent all possible transitions within SPs. A step of synchronization may be necessary for both approaches in order to be able to merge all SPs. Probabilistic analysis have been most of the time used for the fusion, but Multiple Sequence Alignment has also been tested.

Display

Once data are acquired and the model is designed it is generally useful to have a visual representation of the data to easily explore them and to illustrate results. However, complex data structures sometimes prevent straightforward visualisation. High-level tasks recordings of SPMs can be displayed according to two types of visualizations: the temporal and the sequential aspect (Neumuth et al., 2006a). The temporal display more focuses on the duration of each action, whereas the sequential display focused on the relation between work steps. Moreover, in the sequential display, one possibility is to create complete exhaustive tree of each possibility of sequence of work steps. Sensor-based recordings are easier to visualize. As it is represented by time-series data, index-plot can be used. The idea of an index-plot is to display the sequence by representing an activity as a rectangle of specific color for each value, and a width proportional to its duration. Sequence of information can be easily visualized and a quick visual comparison can be performed.

II.2.e. Clinical applications

Clinical applications that are aimed by the analysis and the modelling of surgical procedures are covering multiple surgical specialities, issues and challenges. Five major applications have particularly been of increased attention: 1) evaluation of surgical tools/systems/approaches, 2) training and assessment of surgeons 3) optimization of the OR management 4) context-aware systems, and 5) robotic assistance.We first present surgical specialities that are covered by these systems, and the five main applications are then detailed. A last subsection will permit to present other potential applications.

Surgical speciality

SPMs have been applied to many surgical specialities, but Minimally Invasive Surgery (MIS), including endoscopic and laparoscopic procedures and neurosurgical procedures have been preferred. Within laparoscopic and endoscopic procedures, Cholecystectomy and Functional Endoscopic Sinus Surgery (FESS) surgeries have been widely studied. Eye surgery (Neumuth et al., 2006b, 2011a, 2011b), maxillofacial surgery (Münchenberg et al., 2001), trauma surgery (Agarwal et al., 2007; Bhatia et al., 2007; Xiao et al., 2005), dental implant surgery (Katic et al., 2010), urological surgery (Meng et al., 2004) and otorhinolaryngology (ORL) surgery (Neumuth et al., 2006b) have also been tested. In general, systems were specific to a surgical speciality or even a particular surgery, but a few papers describe more generic surgical systems.

Application

Evaluation of tools/surgical approach/systems: The evaluation of surgical tools or systems has been the first application that has been aimed by research laboratories, on surgeons' demand. Analysis methods that are used in such cases are the comparison and classification methods that allow highlighting a specific parameter such as about surgical tool use, surgical approaches or surgical systems.

Training and assessment of surgeons: All junior surgeons are currently learning with the teaching help of senior surgeons. This is a very time-consuming, interactive and subjective task. From the other hand, there is a growing pressure on surgeons to demonstrate their competences. The need of new automatic training systems with tools for surgeons' evaluation has motivated extensive research into the objective assessment of surgical skills (Hager et al., 2006, Rosen et al., 2001). It would allow surgeons to benefit from constructive feedback, and to learn from their mistakes. Surgical skill can be assessed based on five factors: knowledge, decision making, technical skills, communication skills and leadership skills. From these five factors, technical skills, and especially the dexterity, are vital and based on the surgeon's experience. Historically, this point has been the hardest aspect of assessment to be quantified. It is indeed very subjective because generally performed with questionnaire or human-based observations. Traditional approaches for the assessment of surgical performance rely on prior classification of surgeon technical skills. With automatic techniques, and using a coherent methodology for describing activities, surgical tasks are scored for both precision and speed of performance and are not biased by humans. The ability to precisely recognize simple gesture is a very powerful tool to automate surgical assessment and surgical training. Similar methods can also be employed for training and assessing other members of the surgical team. For a complete discussion on motivations of objective skill evaluation, one can refer to Reiley et al. (2010).

Optimization of the OR management: The need for perioperative surgical workflow optimization has recently emerged, especially regarding the specifications of the OR of the future (Cleary et al., 2005). With the increase number of CAS systems and new technologies, being able to manage and coordinate correctly all these systems is becoming vital. The optimization of the use of physical and human resources required in an OR suite can reduce efforts and therefore improve patient outcomes, reduce hospital's costs and increase efficiency. Moreover, being able to identify different phases within the OR could be useful to know how to assign staff, prepare patient or prioritize OR clean-ups. Additionally, there are some adverse events to take into account. It can be long-time surgical interventions or urgencies that require the use of OR without prior planning. That's why schedule OR has been of increased attention recently (Dexter et al., 2004; Hu et al., 2006) for better management of OR equipments and to facilitate effective allocation of human and material resources.

Context-aware systems: Many CAS systems, such as Augmented Reality (AR) systems or new imaging protocols, have been recently developed and integrated in the OR, but they are used only for a short period of time, and the visualization of additional information strongly depends of the current state of the intervention. Moreover, surgeons have to deal with adverse events during operations, coming from the patient itself but also from the operation management. The idea is to be aware of the difficulties and better handle risks situations. Variations of live signals can be used to warn the surgical staff for anomaly detection. These assistance systems would consequently be of great use for helping, supporting decision making and better managing all CAS systems.

Robotic assistance: Many researches have demonstrated the importance of robots for assistance in surgery, and particularly using SPMs (Ko et al., 2007; Kragic and Hager, 2003; Münchenberg et al., 2001; Miyawaki et al., 2005). They play a vital role in improving accuracy as well as time efficiency in surgical procedures. Two families of robots have been introduced for intra-operative assistance: the semi-active and the active robots. Semi-active robots are making the link between surgeon and patient. Surgeons are performing their tasks outside the OR using the robot that is reproducing surgeon's gesture on the patient. These types of robots are used for specific tasks only such as biopsy or endoscopy for MIS. Active robots are used directly in the OR for replacing the surgeon in certain tasks. Both types of robots could take the benefit of SPMs for supporting these tasks using pre-defined models. The use of robotic assistance also aims at compensating the lack of human resources in many hospitals, and in particular the lack of nurses. The new generation of robots that are currently under testing are situating the intervention progress by automatically acquiring data from the surgical environment and creating SPMs for replacing certain actors of the surgery.

Two other applications that are often implicit in multiple publications are the automatic generation of surgery reports and the help for pre-operative planning.

Surgical reports are papers or informatics files that are generated post-operatively by the surgeon for documenting surgical procedures. Procedures are described as a succession of actions and steps that are manually included into a "log-file" of the surgery for further filing. This step of the procedure is very tedious and time-consuming. The idea of automating this process is to automatically extract as much information as possible from the surgery with the help of multiple sensors, for creating pre-filled reports (Coles and Slavin, 1976). All studies that retrieved information from the OR, regardless to their level of granularity have potentially the possibility of automatically creating pre-filled reports from extracted information.

For helping the pre-operative planning, the recognition of the intervention progress could motivate post-operative discussions between experts in order to better plan next surgeries. The objective is to better anticipate adverse events and possible problems during surgery by using formalized knowledge acquired by previous intervention and also by having an idea of all SPs possibilities. Aggregation and fusion techniques may be helpful in such cases for creating gSPMs.

II.2.f. Validation - Evaluation

We distinguished validation, defined as studying if the system or method is actually doing what it is intended to do, from evaluation, defined as the study of the added value of a system or a method. Each aspect of the SPM methodology is subject to validation. The design of a validation study includes 1) the specification of a validation objective, 2) the definition of input parameters, 2) the computation or estimation of a reference (validation data sets) against which the results of the method to be validated will be compared, 3) the definition of validation metrics that will quantify the comparison, and 4) the operator using the system.

Mainly two principal aspects have been validated; the data acquisition process and the modelling phase. Validation data sets consisted in fully simulated data from computers, data coming from simulated OR, from phantoms or real data directly from surgical interventions and patients. Computer simulations are one possibility of validation data that are easy to create, process and analyse, but are very far from the clinical reality. Similarly, virtual environment (simulated OR) are also quite far from the reality. While both approaches allow a real flexibility for developing new studies, it remains very

difficult to realistically model a surgical environment, like haptic feedbacks or visual effects of the surgeon/patient interaction. Moreover, even if the simulation is close to the reality, the human factor is missing and could bias applications that are intended to be used in real OR environments. The third possibility is to use real surgery devices on phantoms instead of humans. Even if the environment is closer to the reality than complete virtual environments, it remains a part of the procedure that is not realistic. The validation strategies generally consisted in, leave-one-out or k-fold cross-validation approaches. The comparison metrics were the recognition rate (accuracy), the reproducibility, the specificity and the sensitivity.

Few evaluation studies have been conducted and reported in the literature. Some papers indirectly show the added value of SPM approach through its use for comparison of populations of surgical cases performed with different systems or by surgeons with different surgical expertise. For the few papers that are evaluating their systems, same elements than the validation part can be defined.

II.2.g. List of publications

We propose an exhaustive list of publications (**Table 3**) that have been reviewed according to the selection process of **Figure 3** and classify according to diagram of **Figure 5**.

II.3. Scope of the materials

In this section, we explain why we didn't include some publications into our review.

From the beginning of the 90s, many clinical studies were published using the principle of time-motion analysis. Time was the first information chosen by teams to evaluate surgical systems, tools, approaches or assess surgeons. Publications covering time-motion analysis are very close to the papers that are cited here from the data acquisition aspect. Indeed, they used off-line observer-based recording from videos (installed in the OR, on the surgeon, or in the operating field) for acquiring sequences of phases/steps/activities that are then processed trough statistical analysis. However, these papers are not methodological papers, restraining the analysis to statistical computations of time or number of occurrences. They are also always published in clinical journals, which make their impact in term of methodology for our review low. We therefore didn't include them into the review, but some major examples of publications are listed here: Weinger et al., 1994; den Boer et al., 2001; Sjoerdsma et al., 2000; Darzi and Mackay, 2002; Bann et al., 2003; Dosis et al., 2004; Mehta et al., 2001; Malik et al., 2003; Cao et al., 1999; Claus et al., 1995; Payandeh et al., 2002. A classification of their data acquisition techniques and modelling is proposed here on **Table 2**.

	Data acquisition				Modelling		
	Granularity level	**Operator +/- body part**	**Moment of acquisition**	**Method for recording**	**Granularity level**	**Operator +/- body part**	**Formalization**
Time-Motion analysis	Steps/Activities/ Motions	Surgeon	Intra	Observer-based recording from video (off-line)	Steps/Activities/ Motions	Surgeon	Hierarchical decomposition

Table 2 - Classification of time-motion analysis publications, for the data acquisition and the modelling component.

Papers	Data acquisition				Modeling			Analysis				Application		Validation			
	Granularity level	Operator +/- body part	Moment of acquisition	Method for recording	Granularity level	Operator +/- body part	Formalization/Representation Expert knowledge	Type of approach	From data to Model Algorithms	Aggregation – Registration Methods	Comparison – Classification Methods	Surgical speciality	Application	Validation Objective	Validation Dataset	Validation Metric	Validation Operator
Agarwal et al. (2007)	Low-level (presence/absence, vital signs, audio)	Surgical staff/tools Patient Surgical staff	Peri	RFID tags (on-line) Patient monitoring systems (on-line) Audio recording systems (on-line)	Activities	Surgical staff Surgeon Patient	Sequential list	Bottom-up	Statistical analysis	X	X	Traumatologic surgery	Context-aware systems	Analysis	Phantom	Recognition rate	X
Ahmadi et al. (2007)	Low-level (presence/absence)	Surgical tools	Post	Observer-based recording from video (off-line)	Phases	Surgeon	Sequential list	Bottom-up	DTW	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Clinical data	Recognition rate	Surgeon
Ahmadi et al. (2009)	Low-level (trajectories)	Surgeon	Post Intra	Observer-based recording from video (off-line) Tracking devices (on-line)	Motions	Surgeon	Unsequential list	Bottom-up	Motif discovery	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Clinical data	Motif discovery indices	Surgeon
Bhatia et al. (2007)	Low-level (video)	Surgical staff	Intra	Video-based recording (on-line)	Surgical procedure	Surgical staff/patient	State-transition diagram	Bottom-up	SVM + HMM	X	X	Traumatologic surgery	Optimization OR management	Analysis	Clinical data	Recognition rate	Surgical staff/ patient
Blum et al. (2008)	Low-level (video)	Surgeon	Intra	Video-based recording (on-line)	Steps	Surgeon	State-transition diagram	Bottom-up	HMM	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Clinical data	Recognition rate	Surgeon
Bouarfa et al. (2010)	Low-level (presence/absence)	Surgical tools	Post	Observer-based recording from video (off-line)	Steps	Surgeon	State-transition diagram	Bottom-up	HMM	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Simulation	Recognition rate	Surgeon
Burgert et al. (2006)	Steps/activities	Surgeon	Pre	Manual collect of information (off-line)	Steps/activities	Surgeon	Heavy-weight ontology	Top-down	Description logic	X	X	Neurosurgery	Context-aware systems	X	X	X	X
Fischer et al. (2005)	Steps	Surgical staff	Post	Observer-based recording from video (off-line)	Steps	Surgeon	Heavy-weight ontology	Top-down	Description logic	X	X	Endoscopic surgery – FESS	Evaluation of tools/systems	X	X	X	X
Forestier et al. (2011)	Activities	Surgeon	Intra	Observer-based recording (on-line)	Activities	Surgeon	Hierarchical decomposition	Same level	X	X	DTW + HAC	Neurosurgery – Lumbar discectomies	Training/assessment of surgeons	Analysis	Clinical data	Recognition rate (clustering)	Surgeon
Houliston et al. (2011)	Low-level (positions)	Anaesthesist	Intra	RFID tags (on-line)	Activities	Anaesthesist	Sequential list	Bottom-up	Neural network	X	X	Unidentified	Evaluation of tools/systems	Analysis	Clinical data	Recognition rate	Anaesthesist
Hu et al. (2006)	Low-level (video, vital signs)	Surgical staff/ Patient	Intra	Video-based recording (on-line) Patient monitoring systems (on-line)	Surgical procedure	Surgical staff/patient	Sequential list	Bottom-up	Signal processing	X	X	Unidentified	Optimization OR management	Analysis	Clinical data	Recognition rate	Surgical staff/patient
Ibbotson et al. (1999)	Low-level (trajectories)	Surgeon (eyes)	Post	Observer-based recording from video (off-line)	Phases/Steps/ activities	Surgeon	Hierarchical decomposition	X	X	X	Statistical comparisons	Endoscopic/Laparoscopic surgery	Evaluation of tools/systems	Analysis	Clinical data	Time/ Frequence comparisons	Surgeon
James et al. (2007)	Low-level (trajectories,video)	Surgeon (eyes) Surgeon	Intra	Tracking devices (on-line) Video-based recording (on-line)	Phases	Surgeon	Sequential list	Bottom-up	Computer vision + neural network	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Clinical data	Recognition rate	Surgeon
Jannin et al. (2003)	Low-level (images)	Patient	Pre	Manual collect of information (off-line)	Steps	Surgeon	UML class diagram	Top-down	Model instantiation	X	X	Neurosurgery	Context-aware systems	Analysis	Clinical data	Semantic validation	Surgeon
Jannin et al. (2007)	Low-level (images)	Patient	Pre	Manual collect of information (off-line)	Steps	Surgeon	UML class diagram	Top-down	Clustering/Decision tree	X	X	Neurosurgery	Context-aware systems	Analysis	Clinical data	Recognition rate	Surgeon
Katic et al. (2010)	Low-level (positions)	Surgical tools Patient	Intra	Tracking devices (on-line)	Phases	Surgeon	Heavy-weight ontology	Top-down Bottom-up	Description logic	X	X	Dental implant surgery	Context-aware systems	Analysis	Phantom	Recognition rate (situation recognition)	Surgeon
Klank et al. (2008)	Low-level (video)	Surgeon	Intra	Video-based recording (on-line)	Phases	Surgeon	Sequential list	Bottom-up	Computer vision + SVM	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Clinical data	Recognition rate	Surgeon
Ko et al. (2007)	Low-level (positions)	Surgical tools	Intra	Robot-supported recording (on-line)	Steps	Surgeon	State-transition diagram Step/tool/camera view model	Top-down Bottom-up	Sequential analysis	X	X	Laparoscopic surgery – Cholecystectomy	Robotic assistance	X	X	X	X
Kragic and Hager (2003)	Motions	Surgical tools	Intra	Robot-supported recording (on-line)	Activities	Surgeon	State-transition diagram + XML schema	Top-down Bottom-up	Sequential analysis	X	X	Endoscopic/Laparoscopic surgery	Robotic assistance	X	X	X	X
Lemke et al. (2004)	Steps	Surgical staff	Post	Observer-based recording from video (off-line)	Steps	Surgical staff	2D graph Hierarchical representation	Top-down	Description logic	X	X	Neurosurgery Maxillofacial surgery	Evaluation of tools/systems	X	X	X	X
Lin et al. (2006)	Low-level (trajectories)	Surgeon	Intra	Robot-supported recording (on-line)	Motions	Surgeon	Unsequential list	Bottom-up	LDA + Bayes classifier	X	X	Endoscopic/Laparoscopic surgery	Training/assessment of surgeons	Analysis	Phantom	Recognition rate	Surgeon
Lo et al. (2003)	Low-level (video)	Surgeon	Intra	Video-based recording (on-line)	Phases	Surgeon	State-transition diagram	Bottom-up	Computer vision + Bayesian network	X	X	Endoscopic/Laparoscopic surgery	Training/assessment of surgeons	Analysis	Clinical data	Recognition rate	Surgeon

MacKenzie et al. (2001)	Phases/Steps/Activities/Motions	Surgeon Surgical tools	Post	Observer-based recording from video (off-line)	Phases/Steps/ Activities/ Motions	Surgeon Surgical tools	Hierarchical decomposition	Same level	X	Probabilistic analysis	X	Endoscopic/Laparoscopic surgery	Evaluation of tools/systems	X	X	X	X
Meng et al. (2004)	Activities	Surgeon	Pre	Manual collect of information (off-line)	Activities	Surgeon	UML class diagram XML schema	Same level	X	Multiple Sequence Alignment	X	Urological surgery	Evaluation of tools/systems	Analysis	Clinical data	Recognition rate	Surgeon
Miyawaki et al. (2005)	Low-level (trajectories)	Surgeon/Nurse (forehead,wrist, elbow,shoulder)	Post	Tracking devices (off-line)	Activities Steps	Nurse	State-transition diagram	Bottom-up	Model checking	X	X	Endoscopic surgery – Lung resection	Robotic assistance	X	X	X	X
Malarme et al. (2010)	Low-level data	Surgeon	Post	Observer-based recording from video (off-line)	Steps	Surgeon	Heavy-weight ontology	Top-down Bottom-up	Temporal rescaling method Inference engine	X	X	Neurosurgery - Spine	Context-aware systems	Analysis	Simulated OR	Recognition rate (situation recognition)	Surgeon
Münchenberg et al. (2001)	Low-level (images)	Patient	Intra	Robot-supported recording (on-line)	Phases/Steps/ Activities	Surgeon	Hierarchical decomposition	Top-down	Sequential analysis	X	X	Maxillo-facial surgery	Robotic assistance	X	X	X	X
Nara et al. (2011)	Low-level (trajectories)	Surgical staff	Intra	Tracking devices (on-line)	Phases	Surgical staff	Sequential list	Bottom-up	Trajectories Data Mining	X	X	Neurosurgery	Optimization OR management	Analysis	Clinical data	Recognition rate	
Neumuth et al. (2006b)	Activities	Surgeon	Intra	Observer-based recording (on-line)	Activities	Surgeon	UML class diagram XML schema	Same level	X	Probabilistic analysis	X	ORL surgery Neurosurgery Eye surgery	Evaluation of tools/systems	X	X	X	X
Neumuth et al. (2009)	Activities	Surgeon	Intra	Observer-based recording (on-line)	Activities	Surgeon	Hierarchical decomposition	Same level	X	X	X	Endoscopic surgery – FESS	Evaluation of tools/systems	Acquisition	Clinical data	Live Vs Video acquisition comparisons	Surgeon
Neumuth et al. (2011a)	Activities	Surgeon	Intra	Observer-based recording (on-line)	Activities	Surgeon	Hierarchical decomposition	Same level	X	X	Similarity metrics	Eye surgery – Cataract Neurosurgery	Evaluation of tools/systems	Analysis Acquisition	Clinical data	Correlation	Surgeon
Neumuth et al. (2011b)	Activities	Surgeon	Intra	Observer-based recording (on-line)	Activities	Surgeon	Hierarchical decomposition	Same level	X	Statistical analysis	X	Eye surgery – Cataract	Evaluation of tools/systems	Analysis	Clinical data	Time/ occurrence comparisons	Surgeon
Nomm et al. (2008)	Low-level (trajectories)	Surgeon	Post	Tracking devices (off-line)	Motions	Surgeon	Unsequential list	Bottom-up	Statistics + Neural network	X	X	Endoscopic/Laparoscopic surgery	Robotic assistance	Analysis	Clinical data	Recognition rate	Surgeon
Padoy et al. (2007)	Low-level (presence/absence)	Surgical tools	Post	Observer-based recording from video (off-line)	Phases	Surgeon	Sequential list	Bottom-up	DTW	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Clinical data	Recognition rate	Surgeon
Padoy et al. (2008)	Low-level (video, presence/absence)	Surgical tools	Intra Post	Video-based recording (on-line) Observer-based recording from video (off-line)	Phases	Surgeon	State-transition diagram	Bottom-up	Computer vision + HMM	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Clinical data	Recognition rate	Surgeon
Padoy et al. (2010)	Low-level (video, presence/absence)	Surgical tools	Intra Post	Video-based recording (on-line) Observer-based recording from video (off-line)	Phases	Surgeon	Sequential list State-transition diagram	Bottom-up	DTW/HMM	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Clinical data	Recognition rate	Surgeon
Qi et al. (2006)	Phases	Surgical staff	Post	Manual collect of information (off-line)	Phases	Surgical staff	State-transition diagram	Top-down	Workflow engine	X	X	Unidentified	Optimization OR management	X	X	X	X
Riffaud et al. (2011)	Activities	Surgeon	Intra	Observer-based recording (on-line)	Activities	Surgeon	Hierarchical decomposition	Same level	X	X	Statistical comparisons	Neurosurgery – Lumbar discectomies	Training/assessment of surgeons	Analysis	Clinical data	Time/ occurrence comparisons	Surgeon
Sandberg et al. (2005)	Surgical procedure	Surgical staff/Patient	Peri	Patient monitoring systems (on-line)	Surgical procedure	Surgical staff/Patient	Sequential list	Same granularity level	X	X	Statistical comparisons	Endoscopic/Laparoscopic surgery	Optimization OR management	Analysis	Simulated OR	Time/occurrence comparisons	Surgical staff/ patient
Speidel et al. (2008)	Low-level (video)	Surgeon	Intra	Video-based recording (on-line)	Activities	Surgeon	Heavy-weight ontology	Top-down Bottom-up	Computer vision + Description logic	X	X	Endoscopic/Laparoscopic surgery	Context-aware systems	X	X	X	X
Sudra et al. (2007)	Low-level (position)	Surgical tools	Intra	Tracking devices (on-line)	Activities	Surgeon	Heavy-weight ontology	Top-down	Description logic	X	X	Laparoscopic surgery – Cholecystectomy	Context-aware systems	Analysis	Phantom	Recognition rate (Reasoning process)	Surgeon
Suzuki et al. (2010)	Low-level (video, audio)	Surgical staff/patient	Intra	Video-based recording (on-line) Audio recording systems (on-line)	Phases	Surgeon	Sequential list	Bottom-up	Signal processing	X	X	Neurosurgery	Context-aware systems	Analysis	Clinical data	Recognition rate	Surgical staff/ patient
Xiao et al. (2005)	Low-level (Vital signs)	Patient	Intra	Patient monitoring systems (on-line)	Surgical procedure	Surgical staff/Patient	Sequential list	Bottom-up	Signal processing	X	X	Traumatologic surgery	Optimization OR management	Analysis	Clinical data	Recognition rate	Surgical staff/ patient
Yoshimitsu et al. (2010)	Low-level (trajectories)	Surgeon/Nurse (forehead,wrist, elbow,shoulder)	Intra	Tracking devices (on-line)	Activities Steps	Nurse	State-transition diagram	Bottom-up	Timed automata	X	X	Endoscopic/Laparoscopic surgery	Robotic assistance	X	X	X	X

Table 3 - Classification of the 43 publications that have been peer-reviewed.

On the other hand, other recent papers using robot-supported recording, like for instance the paper of Hager et al. (2006) or Rosen et al. (2006), are closer to our review in term of methodology than those using time-motion analysis. Fully connected HMMs are used in these papers for classifying hand trajectories for assessing the level of expertise of surgeons. The reason why we didn't consider them is because they don't incorporate any sequential aspect of the surgical processes into their analysis. The model incorporate sequences of activities but that are not constrained. An existing recent review has already been published on the methods for objective surgical skills evaluation (Reiley et al., 2010), which include all papers using trajectories analysis for surgical skills assessment. A non-exhaustive list of these papers is given here: Hager et al. (2006), Rosen et al. (2001, 2002, and 2006), Voros and Hager (2008), Lin et al. (2006). A classification of their data acquisition techniques and modelling is also proposed on **Table 4**.

	Data acquisition				**Modelling**		
	Granularity level	**Operator +/- body part**	**Moment of acquisition**	**Method for recording**	**Granularity level**	**Operator +/- body part**	**Formalization**
Surgical skill evaluation	Motions	Surgeon	Intra-operative	Robot-supported recording (on-line)	Motions	Surgeon	Sequential list of words

Table 4 - Classification of surgical skills evaluation using robot-supported recording publications, for the data acquisition and the modelling component.

Others studies focused on the pre-processing steps before a SPM analysis. Radrich et al. (2008, 2009) presented a system for synchronizing multi-modal information using various signals for surgical workflow analysis. Sielhorst et al. (2005) synchronized 3D movements before the comparison of surgeons' activities. Speidel et al. (2008, 2009) focused on the identification of instruments in MIS, with the goal of improving current intra-operative assistance systems. These studies, as being just pre-processing steps for further SPM analysis, were identically not integrated to the review.

Around the wide field of human and resource modelling for healthcare, some research also don't focused on the modelling of the surgical process, but on hospital systems (Wendt et al., 2003; Winter et al., 2003), hospital data (Maruster et al., 2001; Rosenbloom et al., 2006), or medical process in the hospital focusing on surgical workers activities (Favela et al., 2007; Sanchez et al., 2008). Other research focused on the modelling of the environment of the OR (inside and outside) but without looking at the surgery itself (Riley and Manias, 2005; Sandberg et al., 2005; Archer and Macario, 2006). Purposes are multiple but every project has a unique objective, which is the improvement of the quality of patient care along with a superior medical safety by studying flows or activities. Also, from an anaesthetist point of view, works have been made by looking at the ergonomic and organisation inside the OR (Seim et al., 2005; Schleppers and Bender, 2003; Decker and Bauer, 2003; Gehbard and Brinkmann, 2006). In spite of this consequent number of works for healthcare systems, most interest has been given last few years for the understanding and the comprehension of the OR, and specifically of the surgery. These studies are all not focusing on process of surgery and were therefore not included in the study.

II.4. Discussion

II.4.h. Modelling

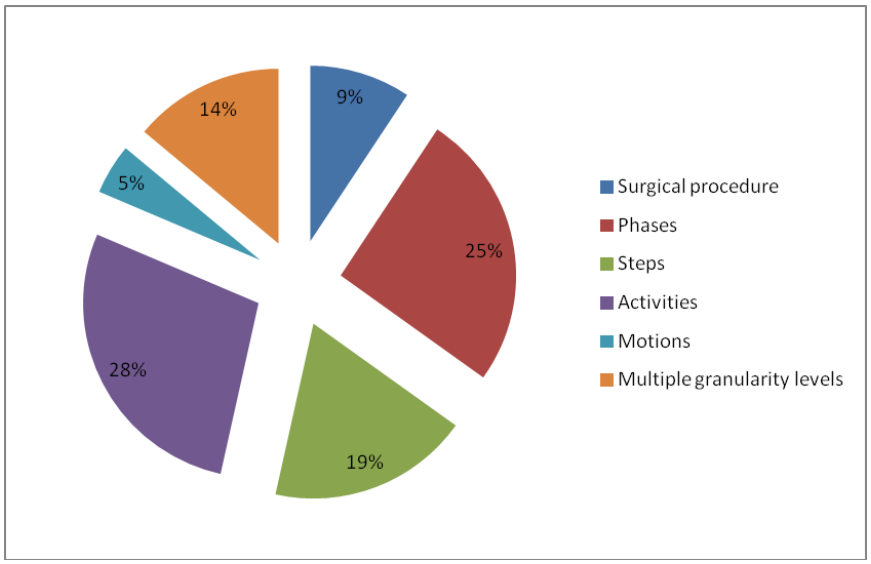


Figure 8 - Repartition of granularity levels of the modelling.

As we can see from **Figure 8**, all granularity levels have been studied, with a particular focus on steps and activities. Moreover, a consequent number of these studies are using multiple granularity levels in their modelling. This type of approach seems to be required for creating global SPMs integrating all aspects of the surgical procedure, and future studies on SPMs should certainly based their modelling aspects on this hypothesis.

From the methods used for formalization, XML schema, which is a light-weighted ontology, defines a grammar that characterizes the structure of a document or the type of data used. They haven't class concept and they also aren't totally dynamic. Actually, XML schema can be a solution for describing SPM at a high level of granularity, to structure data with a well-defined grammar, but they don't respect important concepts such as the classes or the organization into a hierarchy. As XML schema, UML class diagram doesn't allow defining unique and uniform entities. Both approaches seem to be less adapted to the formalization of a surgical context than heavy-weight ontologies. These latter ones allow specifying that two elements correspond to the same unit. In opposition to taxonomies that define classes and relation between these classes, ontology allows defining inference rules. Jannin et al. (2003) proposed a model based on pre and post-operative acquisition of data, including interview of surgeons. The type of surgical procedure, steps and actions were extracted and permit the creation of the model. Additionally, information related to images were linked to classes. Lemke et al. (2004) first defined a surgical ontology as a formal terminology for a hierarchy of concepts and their relationship in the specialized clinical context of surgical procedures and actions. Later, Burgert et al. (2006) proposed an explicit and formal description in an upper-level-ontology

based on General Ontological Language (GOL) for representing surgical interventions. These works were the first one introducing heavy-weight ontologies in the context of surgery.

One important message of this aspect of a SPM methodology is that formalization is needed for being able to compare and share studies between different centres. Even though two centers acquire data on the same surgical procedure using the exact same terminology, a heavy-weighted ontology is still needed to be able to use both data in a common study. The observer based acquisition approach requires the selection of terms from a list to describe the procedure. Two observers may use the same word for two different meanings. Ontology would help making explicit relationships between these two meanings. The more formalization will be used into the modelling, the more semantic will be considered and the more sharable will be the SPM. A heavy and rich formalization is therefore the key for future analysis of SPM to tackle all these issues.

II.4.a. Data acquisition

Both observer-based and sensor-based data acquisition approaches present advantages and drawbacks. Within observer-based approaches, data acquisition process can be supported by two levels of knowledge: the activity recording is performed either according to common standards of surgical procedures or according to fixed-protocol created by local experts. In the first case, standards surgical terms and activities are reported for describing the surgery whereas in the second case, the first step consists of building up its own vocabulary. A new terminology is employed and permits a knowledge representation that is proper to the own surgeon's experience and to the specific surgical environment. The related models are most of the time not based on an ontology, and they are thus not an efficient formal representation of the knowledge and are also not easily sharable between centres. Moreover, the major concern of on-line observer-based approach is the necessity of doing manual labelling that makes the system not automatic and time-consuming. The necessity of having one person in the OR for recording only, whom is often an expert surgeon for reliable information labelling, is also a bias of this approach. At the same time, it's the best way for recording finer details and capture high semantic level, which makes this technique advantageous compared to sensor-based approaches that don’t acquire data at this level of semantic.

Sensor-based approaches are now more and more adopted. For motions detection using tracking systems, the main drawback is that it relies on tools only and motions may not be efficiently detected with rare movements. Compared to other data acquisition techniques, analyses of videos would permit not only to avoid the installation of additional materials in the OR, but also to have a source of information that has not to be controlled by human. For instance, acquiring information from the endoscopic view is very promising for higher level information recognition. Videos are a very rich source of information, as demonstrated on laparoscopy by Speidel et al. (2008). Using image-based analysis, it is possible to acquire relevant information on surgery without disturbing the intervention course. Unfortunately, current image-based algorithms, even with progress in computer vision, don’t allow to completely capturing the well-known semantic gap, in which low-level visual features cannot correctly represent the high-level semantic content of images. Using image-based algorithms, users usually do not think in terms of low-level features, resulting in a bad recognition of the high-level semantic content of images. For model of instruments' use, in spite of high detection accuracies, the major concern is that the recording of signals is not automatic when RFID tags are not used. The entire annotation is performed manually, which makes the system not usable in clinical routine. In practice, RFID tags are too much intrusive, and some vital information that could improve the detection rates

are missing, such as the anatomical structures treated. Generally speaking, all type of sensors additionally installed in the OR show promising results for the challenge of workflow recovery, but the main drawback is the modification of the OR set-up and the necessity of managing such new devices. In particular, eye-gaze tracking systems are interesting because it takes into account the perceptual behaviour of the surgeon, but it would demand large modifications during the intervention course not to alter the clinical routine so far.

In conclusion, observer-based approaches have the capacity to cover high granularity levels for describing surgery, from the lower level (time) to the highest one letting the observer taking the responsibility of acquiring semantic information from pre-defined terminologies and ontologies. On the other hand, it is a very time consuming and costly approach, with the necessity to have a surgeon with a certain clinical background in the OR during the whole procedure. Sensor-based approaches haven't this ability to capture information with semantic meanings, but have the advantage to record live signals automatically or semi-automatically, which is less time-consuming.

For now, no papers are covering multiple levels of granularity, which shows the difficulty of combining different data acquisition methods at different granularity levels. Multiple sensors can be used for instance for both capturing videos and positions of instruments, but the combination of observer-based and sensor-based approaches turns out to be very difficult to set up. We see from **Figure 9** that no predominant techniques have been employed.

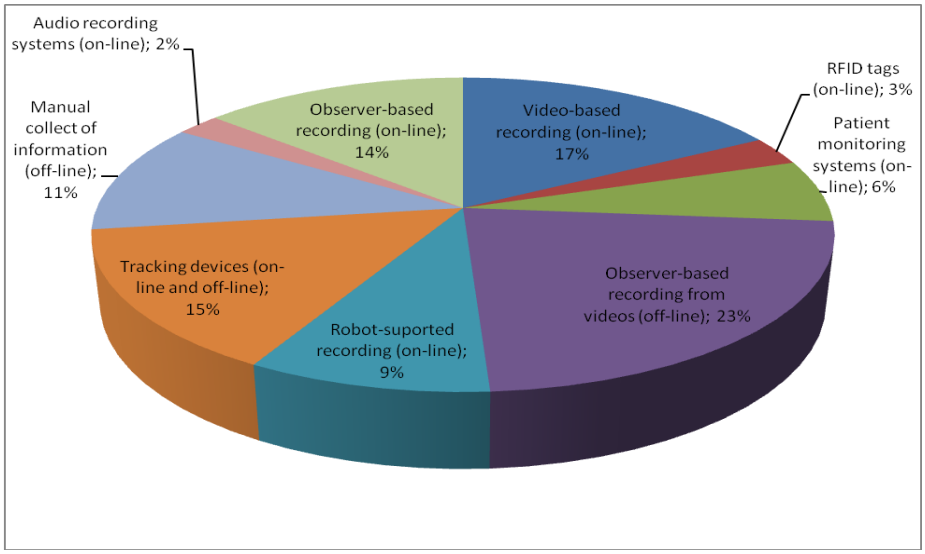


Figure 9 - Repartition of data acquisition techniques

II.4.b. Analysis

The choice of analysis methods that allow going from data to model is vital in current SPM methodology. It allows in the case of bottom-up approaches to bridge the semantic gap between numeric and symbolic data. Based on a preliminary formalization, these methods are all using supervised techniques based on a training stage, except the work of Ahmadi et al. (2007). This type of approach is also the most current one (**Figure 10**). People reports recognition rates from 70% up to 99% but these values are very difficult to compare due to the differences of validation strategies but

also due to the differences of surgical specialities or number and type of data used. The two others approaches (approaches that stay at the same granularity level and top-down approaches), even if they have still not completely shown their interest in the field, are now more and more used and it would be no surprising that the number of publications using both approaches raise in the close future.

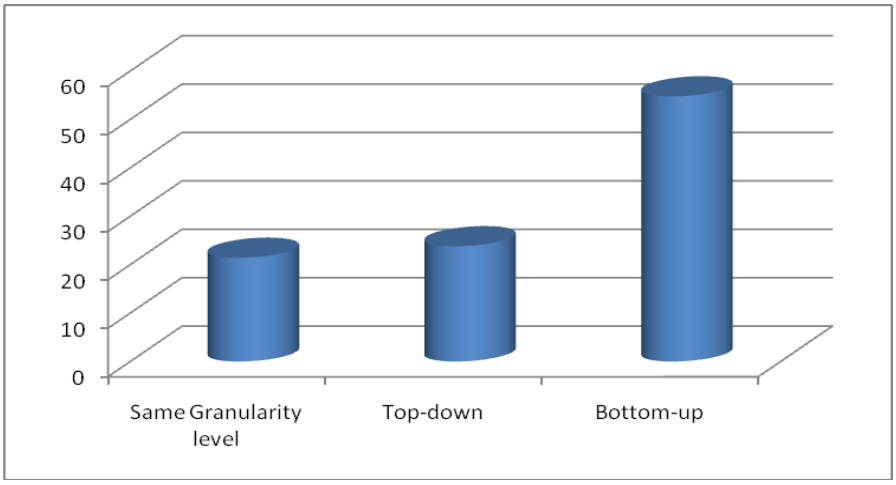


Figure 10 - Repartition of the type of approaches used for "data to model" approaches.

The category of aggregation/fusion analysis method is important because it is a smart way for creating gSPMs that can be used as a supplementary tool for assisting surgeons. It allows creating knowledge models based on an automated SPMs analysis and not on traditional knowledge acquisition methods. The problem of this kind of approach is that it only represents the SPMs that are studied. Even if it clearly seems to be a vital aspect for improving surgery performance, no extensive work has been performed while this type of approach proposes large perspectives in the future. Efforts have therefore to be put here for integrating and automating average models of surgical processes in the clinical routine.

Similar to the previous category of analysis approach, the comparison and classification using surgical processes has not motivated lots of studies yet, but it may also be a direction to take into account. Comparisons of tool uses, surgeons or surgery performance using these kinds of methods allow quantitatively validating and assessing impacts on the surgical procedure.

II.4.c. Applications

We restrained in the diagram potential applications to the 5 most common cited ones in the papers. Additionally, when multiple applications were cited in the papers, we only kept the major clearly identified one. **Figure 11** shows the repartition of applications as well as surgical specialities.

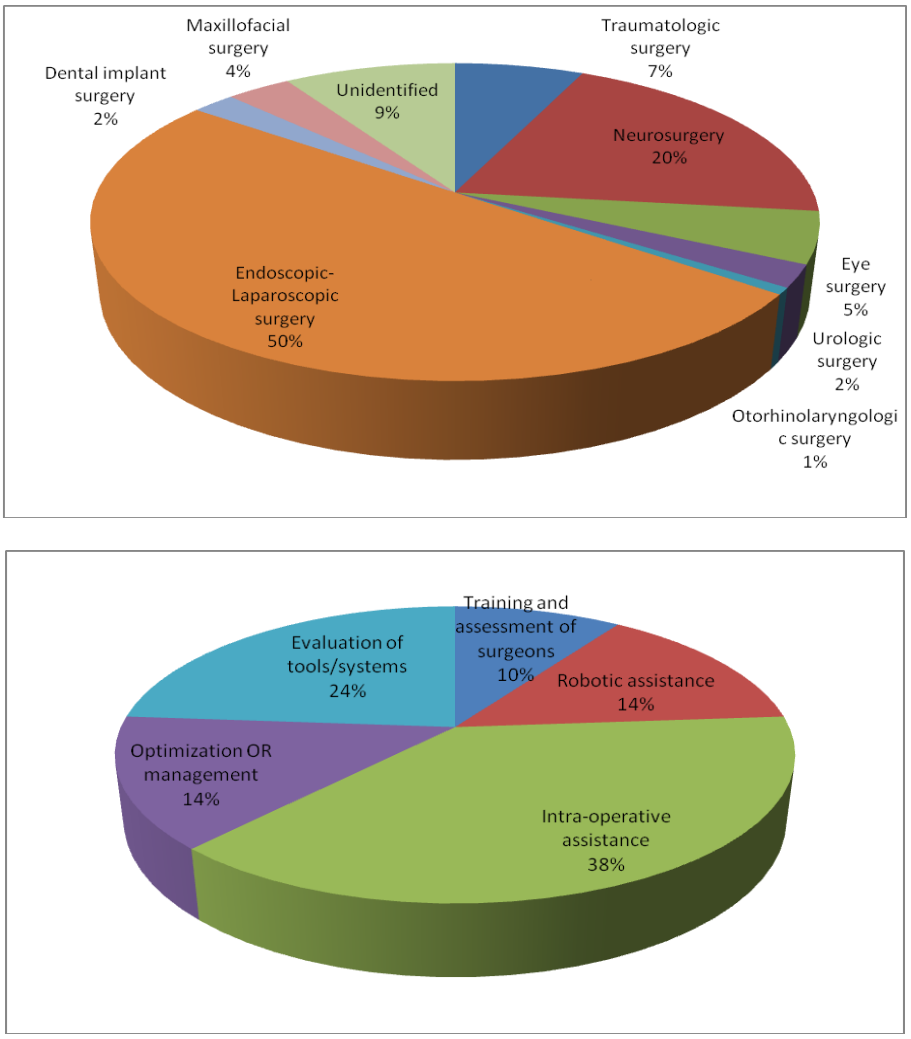


Figure 11 - Repartition of surgical specialities (above) and clinical applications (below).

The majority of the papers describes SPM to be used either in the context of neurosurgery or endoscopy/laparoscopy. This is not surprising, as neurosurgery and MIS have been the most common applications for computer-assisted surgery research. In the case of endoscopic and laparoscopic procedures, surgeries are often very standardized, with a well-defined protocol, widely documented, and inter-patient variabilities remain very low. Data are also easily available for engineers for this surgical speciality. In neurosurgical procedures, data can also be easily acquired. In the case of eye surgeries, new studies are using this surgical speciality because of the very brief procedures.
On the other hand, the repartition of applications is more homogeneous. Even if systems aiming at improving intra-operative assistance are predominant, the four other applications have been seriously and similarly considered. Ahead of the large number of applications cited in publications, we see that SPMs can be useful in the entire surgery timeline, from pre-operative use to post-operative analysis. It can be used in every medical process and adapted to every surgical speciality which shows the potential importance of SPMs.

II.4.d. Validation-Evaluation

The majority of the papers performed validation studies (**Figure 12**, left) on the analysis part (69%), while a very few of them validated the acquisition step (4%). It also remains some studies that don't validate their systems at all (27%). When used, validations studies were performed (**Figure 12**, right) using clinical data in majority (78%). Few of studies are using phantoms, simulated OR or computer simulations.

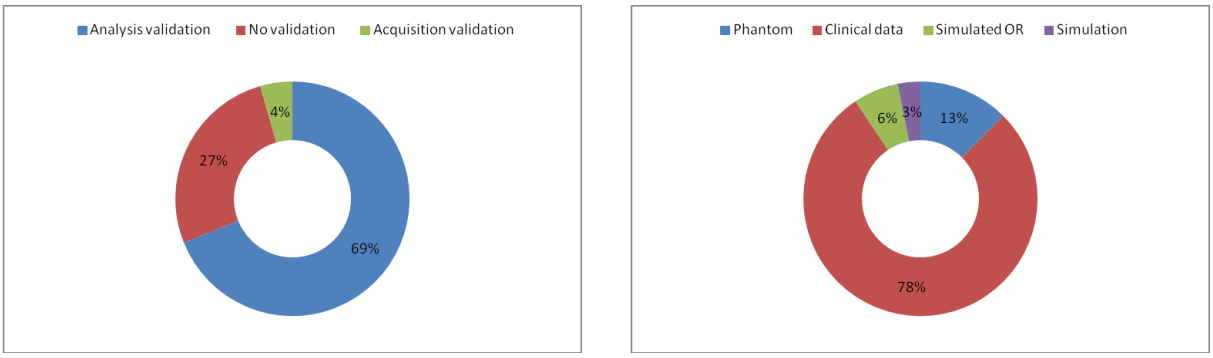


Figure 12 - Repartition of the types of validation (left) and types of validation data (right).

From the 43 publications that were peer-reviewed, only three of them performed evaluation studies. **Table 5** shows the different elements of their evaluation studies.

				Evaluation	
	System evaluated	**Validation objective (Medical context)**	**Dataset**	**Metric**	**Operator**
Katic et al. (2010)	Context-aware augmented reality system	Drilling planned implant	Phantom	Medical usability (questionnaire) Implant position comparison	Surgeon
Ko et al. (2007)	System for intelligent interaction scheme with a robot	Porcine Cholecystectomy	Clinical data	Number of voice command	Surgeon
Yoshimitsu et al. (2010)	Scrub nurse robot	Endoscopic surgery	Clinical data	Instrument targeting time	Nurse

Table 5 - Classification of the 3 publications performing evaluation studies

However, no validation combined to evaluation has been conducted at the same time. There different results on the validation part shows that researches on the field, while being under considerable development, have not been introduced in the clinical routine so far and work remains to be done on the validation part.

II.4.e. Correlations to other information

The correlation of SPMs with other information, such as patient-specific models, is an important perspective of the domain. Patient-specific models are constructed from pre and post-operative patient data such as clinical data or images (Edwards et al., 1995; Biagioli et al., 2006; Verduijn et al., 2007; Kuhan et al., 2002). Correlation between patient outcomes and pre-operative data can then for instance be conducted with the help of probabilistic models such as Bayesian networks.

One other possibility would be to correlate SPMs with the decision-making process of surgeons during the intervention. The decision making in surgery can be conceptualised by two steps, the assessment and the diagnosis of the situation that must be used to select a specific action. The major aspect of the decision-making is that the decision depends on the level of expertise and tasks demand. Dedicated models can be designed for surgical decision-making support by integrating this aspect. Moreover, correlation between pre and post-operative interviews of surgeons with the intra-operative intervention strategy can allow analysing the decision-making process of surgeons, especially under time pressure, and better understand and anticipate further adverse events (Flin et al., 2007; Jalote-Parmar et al., 2008; Morineau et al., 2009).

II.5. Conclusion and problematic of the thesis

Following the growing need of a new generation of CAS systems for the OR, new techniques have emerged based on the modelling of surgical processes. Research studies have been performed toward the development of sophisticated techniques for optimizing, understanding and better managing surgeries and the OR environment based on SPMs. We presented in this Chapter a methodological review on the creation and the analysis of SPMs, focusing on works that model the procedural approach. For structuring the review we proposed a diagram that classifies existing works based on the definition of 5 major aspects of the SPM creation methodology: the acquisition, the modelling, the analysis, the application and the validation/evaluation. Using this classification we presented the existing literature and discussed the different methods and approaches followed by the community. One of the conclusions of this methodological review is that SPMs created by these different approaches may have a large impact in future surgery innovations, whatsoever in planning or intra-operative purposes. The main message to remember around analysis approaches is that no papers currently combined these approaches within one SPM. As no experiments were already performed it is hard to evaluate the impact of such combinations but we could imagine that the use of average models in the case of top-down or bottom-up approaches, or the use of classification techniques combined with average models could be benefit in further studies. Even if the clinical applications are multiple, lot of works remains to be done in the domain, especially on the development of aggregation/fusion and classification analysis methods. Evaluation studies are also missing in the majority of the papers, which shows the relative novelty of the domain that has to be further investigated and developed.

From this literature review, the problematic of the thesis was centred on four main aspects. First, the idea was to develop new methods and tools for the detection of low-level surgical tasks (i.e. the sequence of activities in a surgical procedure) and high-level surgical tasks (i.e. the surgical phases) in the OR trough a bottom-up approach. Second, following the creation of new sensor-based systems using videos in the OR, the idea was to use microscope videos data as the only source of information of the recognition systems. This technique allowed automating the data acquisition process. Third, with this approach, one objective was to cover multiple granularity levels and, finally, to introduce a strong semantic to the models.

References

- ✓ Agarwal S, Joshi A, Finin T, Yesha Y, Ganous T. A pervasive computing system for the operating room of the future. Mobile Networks and Applications. 2007; 12(2,3): 215-28.
- ✓ Ahmadi A, Sielhorst T, Stauder R, Horn M, Feussner H, Navab N. Recovery of surgical workflow without explicit models. Proc MICCAI, Berlin: Springer. 2007; 420-8.
- ✓ Ahmadi A, Padoy N, Rybachuk K, Feussner H, Heining SM, Navab, N. Motif discovery in OR sensor data with application to surgical workflow analysis and acitivity detection. M2CAI workshop, MICCAI. 2009.
- ✓ Archer T and Macario A. The drive for operating room efficiency will increase quality of patient care. Curr Opin Anaesthesiol. 2006; 19: 171-6.
- ✓ Bann MS, Khan, Darzi A. Measurement of surgical dexterity using motion analysis of simple bench skills. World J Surg. 2003; 27: 390-4.
- ✓ Biagioli B, Scolletta S, Cevenini G, Barbini E, Giomarelli P, Barbini P. A multivariate Bayesian model for assessing morbidity after coronary artery surgery. Crit Care. 2006; 10(3): R94.
- ✓ Blum T, Padoy N, Feussner H, Navab N. Workflow mining for visualization and analysis of surgeries. Int J Comput Assisted Radiol Surg. 2008; 3(5): 379-86.
- ✓ Bouarfa L, Jonker PP, Dankelman J. Discovery of high-level tasks in the operating room. J Biomed Inform. 2010.
- ✓ Bhatia B, Oates T, Xiao Y, Hu P. Real-time identification of operating room state from video. AAAI.2007; 1761-6.
- ✓ Burgert O, Neumuth T, Lempp F, Mudunuri R, Meixensberger J, Strauß G, Dietz A, Jannin P, Lemke HU. Linking top-level ontologies and surgical workflows. Int J Comput Assisted Radiol Surg. 2006; 1(1): 437-8.
- ✓ Cao CGL, MacKenzie CL, Payandeh S. Task and Motion Analysis in Endoscopic Surgery. Proc ASME Dynamic Systems, 5th Annual Symposium on Haptic Interface for Virtual Environment and Teleoperation. 1996.
- ✓ Claus GP, Sjoerdsma W, Jansen A, Grimbergen CA. Quantitative standardised analysis of advanced laparoscopic surgical procedures. Endosc Surg Allied Technol. 1995; 3: 210-3.
- ✓ Coles EC and Slavin G. An evaluation of automatic coding of surgical pathology reports. J Clin Pathol. 1976; 29(7): 621-6.
- ✓ Darzi A, Mackay S. Skills assessment of surgeons. Surgery. 2002; 131(2): 121-4.
- ✓ Decker K and Bauer M. Ergonomics in the Operating Room. Minim Invasive Ther Allied Technol. 2003; 12(6): 268-77.
- ✓ Den Boer KT, de Wit LT, Davids PHP, Dankelman J, Gouma DJ. Analysis of the quality and efficiency of learning laparoscopic skills. Surg Endosc. 2001; 15: 497-503.
- ✓ Dexter F, Epstein RH, Traub RD, Xiao Y.Making management decisions on the day of surgery based on operating room efficiency and patient waiting times. Anesthesiology. 2004; 101(6): 1444-53.
- ✓ Dosis A, Bello F, Moorthy K, Munz Y, Gillies D, Darzi A. Real-time synchronization of kinematic and video data for the comprehensive assessment of surgical skills. Stud Health Technol Inform. 2004; 98:82-8 Edwards FH, Peterson RF, Bridges C, Ceithaml EL. 1988: Use of a Bayesian statistical model for risk assessment in coronary artery surgery. Updated in 1995. Ann Thorac Surg. 1995; 59(6): 1611-2.
- ✓ Favela J, Tentori M, Castro LA, Gonzalez VM, Moran EB, Martinez-Garcia AI. Activity recognition for context-aware hospital applications: issues and opportunities for the deployment of pervasive networks. Mobile Networks Applications. 2007; 12(2,3): 155-71.

- ✓ Fischer M, Strauss G, Burgert O, Dietz A, Trantakis C, Meixensberger J, Lemke HU. ENT-surgical workflow as an instrument to assess the efficiency of technological developments in medicine. International Congress Series (Comput Assisted Radiol Surg).. 2005; 851-5.
- ✓ Flin R, Youngson G, Yule S. How do surgeons make intraoperative decisions. Qual Saf Health Care. 2007; 16: 235-9.
- ✓ Forestier G, Lalys F, Riffaud L, Trelhu B, Jannin P. Classification of surgical processes using dynamic time warping. J Biomed Inform. 2011; 45: 255-64.
- ✓ Gehbard F and Brinkmann, A. Management of an operating room in a university hospital. Zentralbl Chir. 2006; 131(4): 341-6.
- ✓ Hager G, Vagvolgyi B, Yuh D. Stereoscopic video overlay with deformable registration. Medicine Meets Virtual Reality. 2007.
- ✓ Houliston BR, Parry DT, Merry AF. TADAA: Towards automated detection of anaesthetic activity. Methods of Information in Medicine. 2011; 50(5): 464-71.
- ✓ Hu P, Ho D, MacKenzie CF, Hu H, Martz D, Jacobs J, Voit R, Xiao Y. Advanced Visualization platform for surgical operating room coordination. Distributed video board system. Surg innovation. 2006; 13(2): 129-35.
- ✓ Ibbotson JA, MacKenzie CL, Cao CG, Lomax AJ. Gaze patterns in laparoscopic surgery. Stud Health Technol Inform. 1999; 7: 154-60.
- ✓ Jalote-Parmar A, van Alfen M, Hermans JJ. Workflow Driven User Interface for Radiological System: A Human Factors Approach. Int J Comput Assisted Radiol Surg. 2008.
- ✓ James A, Vieira D, Lo BPL, Darzi A, Yang GZ. Eye-gaze driven surgical workflow segmentation. Proc MICCAI. 2007; 110-7.
- ✓ Jannin P, Raimbault M, Morandi X, Riffaud L, Gibaud B. Model of surgical procedures for multimodal image-guided neurosurgery. Computer Aided Surgery. 2003; 8(2): 98-106.
- ✓ Jannin P Morandi X. Surgical models for computer-assisted neurosurgery. Neuroimage. 2007; 37(3): 783-91.
- ✓ Katic D, Sudra G, Speidel S, Castrillon-Oberndorfer G, Eggers G, Dillman R. Knowledge-based situation interpretation for context-aware augmented reality in dental implant surgery. Proc Medical Imaging Augmented Reality. 2010.
- ✓ Klank U, Padoy N, Feussner H, Navab N. Automatic feature generation in endoscopic images. Int J Comput Assisted Radiol Surg. 2008; 3(3,4): 331-9.
- ✓ Ko SY, Kim J, Lee WJ, Kwon DS. Surgery task model for intelligent interaction between surgeon and laparoscopic assistant robot. Journal of Robotics and Mechatronics. 2007; 8(1): 38-46.
- ✓ Kuhan G, Marshall EC, Abidia AF, Chetter IC, McCollum PT. A Bayesian hierarchical approach to comparative audit for carotid surgery. Eur J Vasc Endovasc Surg. 2002; 24(6): 505-15.
- ✓ Kuhnapfel U, Cakmak HK, Maab H. Endoscopic Surgery Training using Virtual Reality and Deformable Tissue Simulation. Computer and Graphics. 2000; 24: 671-82.
- ✓ Lemke HU, Trantakis C, Köchy K, Müller A, Strauss G, Meixensberger J. Workflow analysis for mechatronic and imaging assistance in head surgery. Int Congress Series. 2004; 1268: 830-5.
- ✓ Lin HC, Shafran I, Yuh D, Hager GD. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Computer Aided Surgery. 2006; 11(5): 220-30.
- ✓ Lo B, Darzi A, Yang G. Episode Classification for the Analysis of Tissue-Instrument Interaction with Multiple Visual Cues. Proc MICCAI. 2003.
- ✓ MacKenzie CL, Ibbotson AJ, Cao CGL, Lomax A. Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. Minim Invasive Ther Allied Technol. 2001; 10(3): 121-8.

- ✓ Malik R, White P, Macewen C. Using human reliability analysis to detect surgical error in endoscopic DCR surgery. Clin Otolaryngol Allied Sci. 2003; 28: 456-60.
- ✓ Maruster L, van der Aalst W, Weijters T, van den Bosch A, Daelemans W. Automatic discovery of workflows models from hospital data. Proc BNAIC. 2001; 183-90.
- ✓ Marvik R, Lango T, Yavuz Y. An experimental operating room project for advanced laparoscopic surgery. Semin Laparosc Surg. 2004; 11: 211-6.
- ✓ Mehta NY, Haluck RS, Frecker MI, Snyder AJ. Sequence and task analysis of instrument use in common laparoscopic procedures. Surgical Endoscopy. 2002; 16(2): 280-5.
- ✓ Meng F, D'Avolio LW, Chen AA, Taira RK, Kangarloo H. Generating models of surgical procedures using UMLS concepts and multiple sequence alignment. Proc AMIA. 2005; 520-524.
- ✓ Meyer MA, Levine WC, Egan MT, Cohen BJ, Spitz G, Garcia P, Chueh H, Sandberg WS. A computerized perioperative data integration and display system. Int J Comput Assisted Radiol Surg. 2007; 2(3,4): 191-202.
- ✓ Miyawaki F, Masamune K, Suzuki S, Yoshimitsu K, Vain J. Scrub nurse and timed-automata-based model for surgery. IEEE Industrial Electronics Trans. 2005; 5(52): 1227-35.
- ✓ Morineau T, Morandi X, Le Moëllic N, Diabira S, Haegelen C, Hénaux PL, Jannin P. Decision making during preoperative surgical planning. Human factors. 2009; 51(1): 66-77.
- ✓ Munchenberg J, Brief J, Raczkowsky J, Wörn H, Hassfeld S, Mühling J. Operation Planning of Robot Supported Surgical Interventions. Int Conf Intelligent Robots Systems. 2001; 547-52.
- ✓ Nara A, Izumi K, Iseki H, Suzuki T, Nambu K, Sakurai Y. Surgical workflow monitoring based on trajectory data mining. New frontiers in Artificial Intelligence. 2011; 6797: 283-91.
- ✓ Neumuth T, Schumann S, Strauss G, Jannin P, Meixensberger J, Dietz A, Lemke HU, Burgert O. Visualization options for surgical workflows. Int J Comput Assisted Radiol Surg. 2006a; 1(1): 438-40.
- ✓ Neumuth T, Durstewitz N, Fischer M, Strauss G, Dietz A, Meixensberger J, Jannin P, Cleary K, Lemke HU, Burgert O. Structured recording of intraoperative surgical workflows. SPIE medical imaging - PACS in Surgery. 2006b: 6145; 61450A.
- ✓ Neumuth T, Trantakis C, Eckhardt F, Dengl M, Meixensberger J, Burgert O. Supporting the analysis of inter-vention courses with surgical process models on the example of fourteen microsurgical lumbar discectomies. Int J Comput Assisted Radiol Surg. 2007; 2(1): 436-8.
- ✓ Neumuth T, Jannin P, Strauss G, Meixensberger J, Burgert O. Validation of Knowledge Acquisition for Surgical Process Models. J AMIA. 2008; 16(1): 72-82.
- ✓ Nomm S, Petlenkov E, Vain J, Belikov J, Miyawaki F, Yoshimitsu K. Recognition of the surgeon's motions during endoscopic operation by statistics based algorithm and neural networks based ANARX models. Proc Int Fed Automatic Control. 2008.
- ✓ Padoy N, Horn M, Feussner H, Berger M, Navab N. Recovery of surgical workflow: a model-based approach. Int J Comput Assisted Radiol Surg. 2007; 2(1): 481-2.
- ✓ Padoy N, Blum T, Feuner H, Berger MO, Navab N. On-line recognition of surgical activity for monitoring in the operating room. Proc Conference on Innovative Applications of Artificial Intelligence. 2008.
- ✓ Padoy N, Blum T, Ahmadi SA, Feussner H, Berger MO, Navab N. Statistical modeling and recognition of surgical workflow. Med Image Anal. 2010; 16(3): 632-41.
- ✓ Payandeh S, Lomax AJ, Dill J, Mackenzie CL, Cao CGL. On Defining Metrics for Assessing Laparoscopic Surgical Skills in a Virtual Training Environment. Stud Health Technol Inform. 2002; 85:334-40.
- ✓ Payne PRO, Mendonca EA, Johnson SB, Starren JB. Conceptual knowledge acquisition in biomedicine: a methodological review. J Biomed Inform. 2007; 40(5): 582-602.

- ✓ Qi J, Jiang Z, Zhang G, Miao R, Su Q. A surgical management information system driven by workflow. IEEE conf on service operations and logistics, and informatics. 2006; 1014-8.
- ✓ Radrich H. Vision-based motion monitoring trough data fusion from a chirurgical multi-camera recording system. Diploma thesis. TUM, Munich. 2008
- ✓ Radrich H, Padoy N, Ahmadi A, Feussner H, Hager G, Burschka D, Knoll A. Synchronized multimodal recording system for laparoscopic minimally invasive surgeries. M2CAI workshop, MICCAI;2009.
- ✓ Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. Surgical endoscopy. 2010.
- ✓ Riedl S. Modern operating room mangement in the workflow of surgery. Spectrum of tasks and challenges of the future. Der Chirurg. 2002; 73: 105-10.
- ✓ Riley R and Manias E. Governing Time in Operating Rooms. J Clin Nurs; 2005:15(5); 546-53
- ✓ Rosen J, Hannaford B, Sinanan M, Solazzo M. Objective Evaluation Of Laparoscopic Surgical Skills Using Hidden Markov Models Based On Haptic Information And Tool/tissue Interactions. Stud Health Technol Inform. 2001; 81: 417-23.
- ✓ Rosen J, Solazzo M, Hannaford B, Sinanan M. Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden markov model. Comput Aided Surg. 2002; 7(1): 49–61.
- ✓ Rosen J, Brown JD, Chang L, Sinanan M, Hannaford B. Generalized Approach for Modeling Minimally Invasive Surgery as a Stochastic Process Using a Discrete Markov Model. IEEE Trans Biomed Engineering. 2006; 53(1): 399-413.
- ✓ Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006; 13(3): 277-88.
- ✓ Sanchez D., Tentori, M., Favela, J. Activity recognition for the smart hospital. IEEE intelligent systems. 2008; 23(2): 50-7.
- ✓ Sandberg WS, Daily B, Egan MT, Stahl JE, Goldman JM, Wiklund RA, Rattner D. Deliberate perioperative systems design improves operating room throughput. Anesthesiology. 2005; 103: 406-18.
- ✓ Satava R, Cuschieri A, Hamdorf J. Metrics for objective assessment. Surg Endosc. 2003; 17(2):220-6.
- ✓ Schleppers A and Bender H. Optimised workflow and organisation – from the point of view of an anaesthesiolo-g department. Minim Invasive Ther Allied Technol.. 2003; 12(6): 278-83.
- ✓ Seim AR, Meyer M, Sandberg WS. Does parallel workflow impact anaesthesia quality. Proc AMIA . 2005; 1053.
- ✓ Sielhorst, T., Blum, T., Navab, N. Synchronizing 3d movements for quantitative comparison and simultaneous visualization of actions. Proc. ISMAR. 2005.
- ✓ Sjoerdsma W, Meijer D, Jansen A, den Boer KT, Grimbergen CA. Comparison of efficiencies of three techniques for colon surgery. J Laparoendosc Adv Surg Tech. 2000; 10(1): 47-53.
- ✓ Speidel S, Sudra G, Senemaud J, Drentschew M, Müller-Stich BP, Gun C, Dillmann R. Recognition of risk situations based on endoscopic instrument tracking and knowledge based situation modeling. Progress Biomed Optics Imaging. 2008; 9(1): 35.
- ✓ Speidel S, Benzko J, Krappe S, Sudra G, Azad P, Müller-Stich BP, Gutt C, Dillmann R. Automatic classification of minimally invasive instruments based on endoscopic image sequences. Progress Biomed Optics Imaging. 2009; 10(1): 37.
- ✓ Sudra G, Speidel S, Fritz D, Möller-Stich BP, Gutt C, Dillmann R. MEDIASSIST: MEDIcal ASSITance for intraoperative skill transfer in minimally invasive surgery using augmented reality. Progress Biomed Optics Imaging. 2007; 8(2).

- ✓ Suzuki T, Sakurai Y, Yoshimitsu K, Nambu K, Muragaki Y, Iseki H. Intraoperative multichannel audio-visual information recording and automatic surgical phase and incident detection. Int Conf IEEE EMBS. 2010; 1190-3.
- ✓ Verduijn M, Rosseel PM, Peek N, de Jonge E, de Mol BA. Prognostic Bayesian networks II: an application in the domain of cardiac surgery. J Biomed Inform. 2007; 40(6): 619-49
- ✓ Voros S and Hager GD. Towards "real-time" tool-tissue interaction detection in robotically assisted laparoscopy. Int Conf IEEE RAS ans EMBS. 2008: 562-7
- ✓ Weinger MB, Herndon OW, Zornow MH, Paulus MP, Gaba DM, Dallen LT. An objective methodology for task analysis and workload assessment in anesthesia providers. Anesthesiology. 1994: 80(1); 77-92.
- ✓ Wendt T, Häber A, Brigl B, Winter A. Modeling hospital information systems (Part 2): using the 3LMG2 tool for modelling patient, record management. Methods Inf Med; 2003: 43(3); 256-67
- ✓ WFMC – Workflow management coalition. Terminology & glossary. Doc number WFMC-T-1011, Issue 3.0. Winchester, UK.1999.
- ✓ Winter A, Brigl B, Wendt T. Modeling Hospital information systems (Part 1): The revised three-layer graph-based meta model 3LGM2. Method Inf Med. 2003; 42(5): 544-51
- ✓ Yoshimitsu K, Masamune K, Iseki H, Fukui Y, Hashimoto D, Miyawaki F. Development of scrub nurse robot (SNR) systems for endoscopic and laparoscopic surgery. Micro-NanoMechatronics and Human Science. 2010; 83-88.

PART II

METHODOLOGY FOR AUTOMATIC RECOGNITION OF HIGH-LEVEL AND LOW-LEVEL TASKS IN THE OR

This part of the manuscript is composed by the core of the research performed during my PhD work. In Chapter II, I first present the two data-sets used for our experiments. Chapter IV describes the first method for detecting surgical tasks from microscope videos. This method is based on an image classification problem within a static approach. It shows the feasibility of the detection using global image features only. Then, in Chapter V, the dynamic aspect is incorporated as well as local image features within a global recognition framework in order to improve the recognition rate. In Chapter V, the methodology is extended to address recognition of tasks at a lower granularity level. The proposed methodology relies in additional knowledge of the surgical procedure. In Chapter VII, I propose a general discussion on the advantages of using microscope videos for the recognition task and possible clinical applications. I finally conclude the thesis in Chapter VIII and open the research area to new perspectives.

Chapter III. Data-sets presentation

All studied algorithms and frameworks were evaluated on two data-sets coming from two different surgical specialities. The first one is a data-set of videos of pituitary surgeries. The second one is a data-set of videos of cataract surgeries. We present both data-sets in the next two subsections as well as the different surgical tasks of each surgical procedure, at both high- and low- granularity levels.

III.1. Dataset 1: Pituitary surgery

III.1.a. Surgical procedure

Neurosurgery is a surgical speciality that concerns all pathologies of the central nervous system. From the high number of tumours in neurosurgery, we choose the pituitary adenoma surgeries (Ezzat et al., 2004) which are tumours that occur in the pituitary gland and represent around ten percent of all intra-cranial tumour removals. Neurosurgeons mostly use a direct trans-nasal approach, where an incision is made in the back wall of the nose. Rarely, a craniotomy is required. In this study, all surgeries were performed according to the first approach. This first dataset included 16 pituitary surgeries (with mean surgical time of 50 +/- 8 min), all performed at the neurosurgical department of the University Hospital of Rennes by three expert surgeons. Videos were recorded using the OPMI Pentero surgical microscope (Carl Zeiss® Medical Systems[1], Germany). The initial video resolution was 768 x 576 pixels at 33 frames per second (fps). Recordings were obtained from nasal incision until compress installation, where the microscope was continuously used.

III.1.b. Surgical phases identification

An expert neurosurgeon was asked to decompose a standard pituitary surgical procedure into a set of sequential ordered phases. The initial consign was to identify a set of phases (high-level tasks) that all have a well-defined surgical objective. Mathematically, a surgical process defined at a high granularity level sp_{HL} can be defined as a sequence of phases belonging to the set of phases PH :

$$sp_{HL} =< ph_1, ph_2, ph_3 ..., ph_{n-1}, ph_n > \;\Big|\; ph_i \in PH \tag{1}$$

with $ph_1 \neq ph_2 \neq ph_3 ... \neq ph_{n-1} \neq ph_n$ and n the number of phases.

Six phases (considered as the high granularity level) were defined. These phases are: 1) nasal incision, 2) nose retractors installation, 3) access to the tumour along with tumour removal, 4) column of nose

[1] http://www.meditec.zeiss.com/

replacement, 5) suturing, and 6) nose compress installation. These six phases were also validated by another expert neurosurgeon (**Figure 13**).

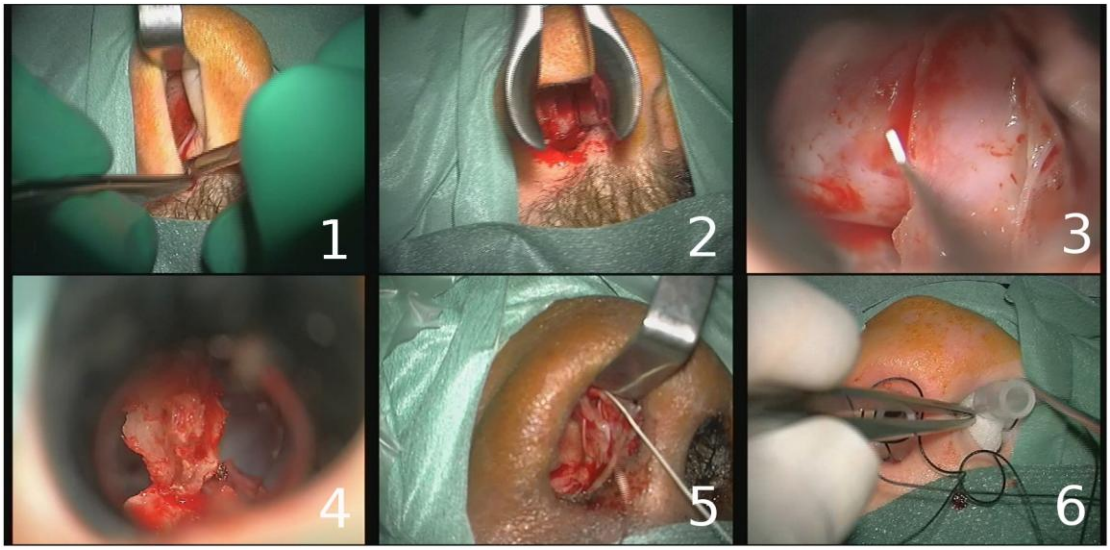


Figure 13 - Example of typical digital microscope images for pituitary surgeries.

III.2. Dataset 2: cataract surgeries

III.2.a. Surgical procedure

The second dataset included cataract surgeries, which is one of the most frequent ophtalmological surgeries in the world. The term cataract refers to the clouding of the normally clear lens of the eye. The corresponding surgery removes the natural lens and inserts an artificial one (namely an IntraOcular Lens, IOL) in order to restore focusing power. This second dataset included 20 cataract surgeries (with mean surgical time of 15 +/- 4 min), all performed at the University Hospital of Munich by two expert surgeons. Videos were recorded using the OPMI Lumera surgical microscope (Carl Zeiss® Medical Systems, Germany) with a resolution of 720 x 576 at 25 fps.

III.2.b. Surgical phases identification

Similar to the definition of the phases for pituitary surgeries, an expert ophthalmological surgeon was asked to decompose a standard cataract surgical procedure into a sequence of phases. Eight surgical phases were defined: 1) preparation of the patient, 2) betadine (antiseptic) injection, 3) access to the anterior chamber through a corneal incision and viscoelastic injection, 4) hydrodissection + capsulorhexis, 5) phacoemulsification, 6) irrigation + cortical aspiration of the remanescent lens, 7) implantation of the artificial IOL, and 8) IOL adjustment + wound sealing. These eight phases were also validated by another expert ophtalmological surgeon (**Figure 14**).

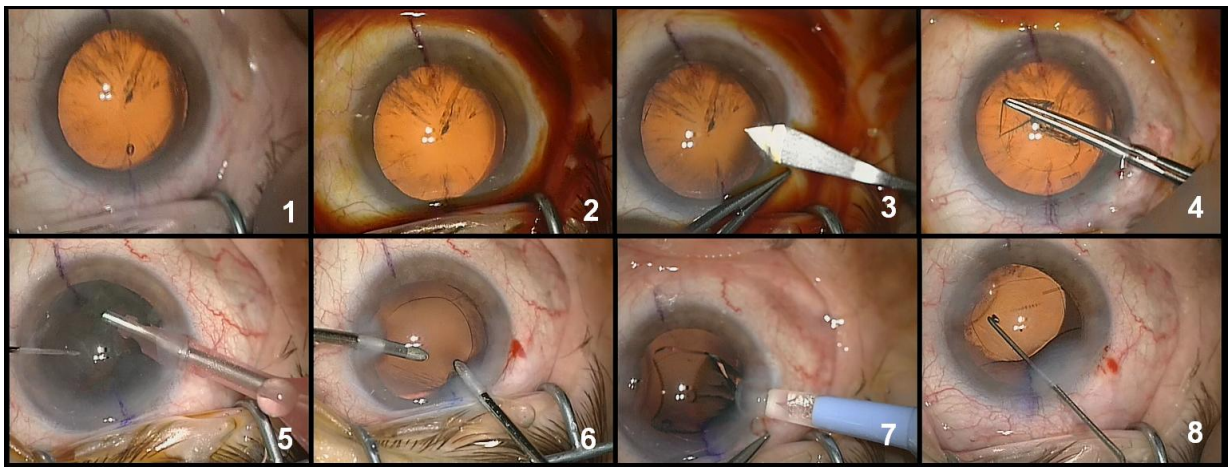


Figure 14 - Example of typical digital microscope images for cataract surgeries.

III.2.c. Surgical activities identification

For identifying the lower granularity level consisting in surgical activities, we relied in a formalism introduced in Neumuth et al. (2006, 2007, and 2008). According to this formalism, an activity ac_i is defined by a triplet:

$$ac_i = < a, i, s > \qquad a \in A, i \in I \ , s \in S, \tag{2}$$

where A is the set of possible actions (e.g. irrigate, cut), I the set of possible instruments (e.g. 1.1 mm knife, micro spatula), and S the set of possible anatomical structures (e.g. cornea, conjunctiva). An example of activity in the context of cataract surgery could be <*aspirate, aspiration cannula, lens*>. The definition domain is thus defined by: $A \times S \times I$. Each activity has also a starting point $start(ac_i)$ and a stopping point $stop(ac_i)$. Note that $start(ac_i) < stop(ac_i)$ and $stop(ac_i) < start(ac_{i+1})$. As the surgeon can hold two surgical tools at the same time, one in each hand, a surgical process defined at a low granularity level sp_{LL} can then be defined as a sequence of activities performed with the right and left hand simultaneously. Each activity belongs to the set of activities AC performed during this dedicated type of surgery:

$$sp_{LL} = \begin{cases} < {ac_1}^L, {ac_2}^L, {ac_3}^L ..., {ac_{n-1}}^L, {ac_{n^L}}^L > \ \middle| \ {ac_i}^L \in AC \\ < {ac_1}^R, {ac_2}^R, {ac_3}^R ..., {ac_{n-1}}^R, {ac_{n^R}}^R > \ \middle| \ {ac_i}^R \in AC \end{cases} \tag{3}$$

with n^L and n^R the number of activities for the left and right hand respectively.

According to this formalization, a cataract terminology was defined by the surgeon. Twelve actions, 13 surgical tools and 6 structures were identified. Using this terminology, 18 activities (as combinations of the three components) were then defined (**Figure 15**). Then, from the possible list of all combinations of activities (i.e. one activity for the left hand and one activity for the right hand), only 24 were kept. In addition to these 24 possible pairs of activities, we defined an activity "background" representing a frame with no activity, i.e. a frame with no tool. We therefore finally obtained set of 25 possible pairs of activities. The videos were post-operatively labelled using the ICCAS editor software (Appendix A) for creating the ground truth.

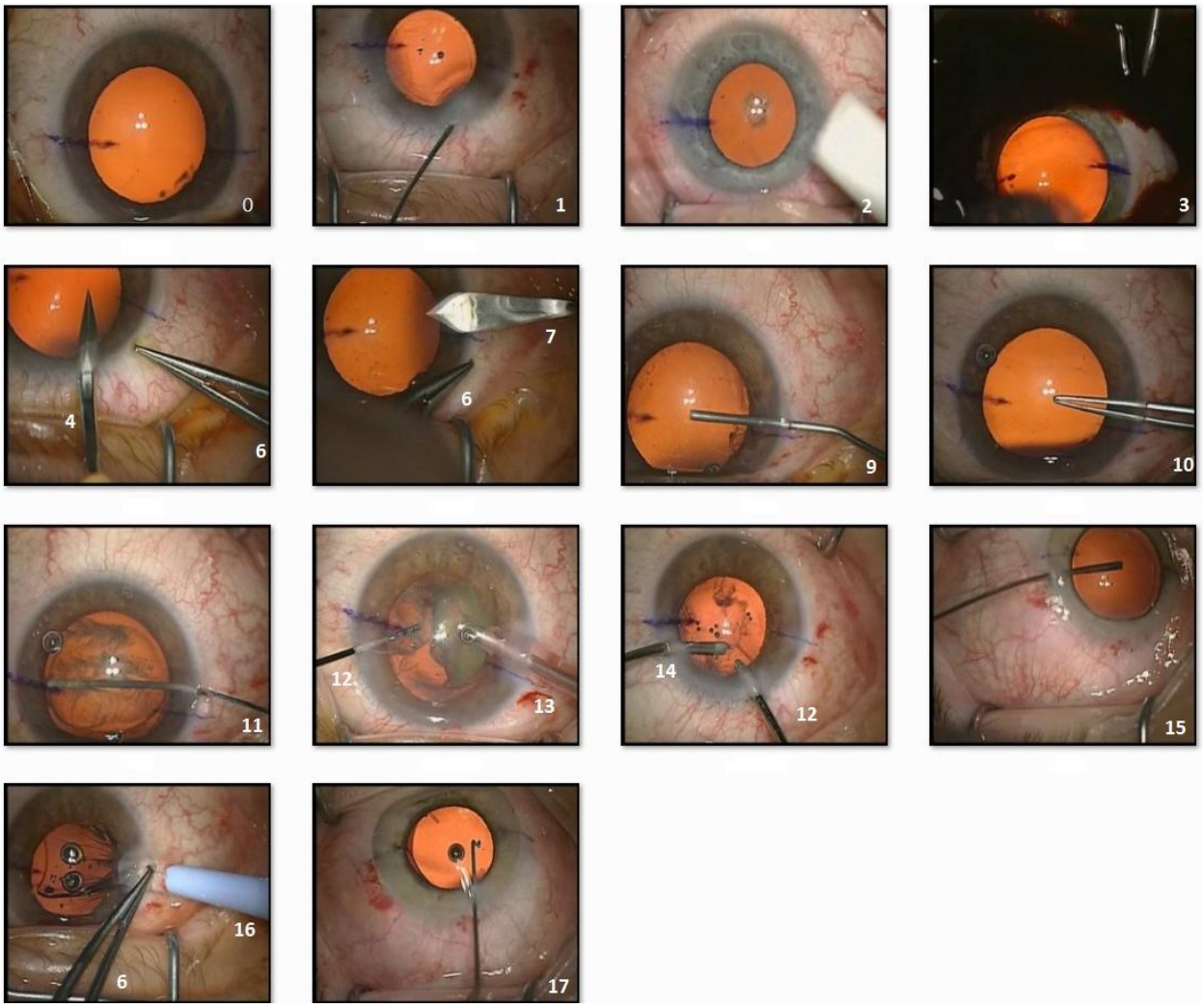


Figure 15 - Example of image frame for each activity.

Table 6 - List of the 18 activities (corresponding to the numbering of **Figure 15**).

Activity	Action	Surgical tool	Structure
n°0	Background	Background	Background
n°1	Wash	Irrigation cannula	Conjunctiva and cornea
n°2	Swab	Swab pagasling	Conjunctiva and cornea
n°3	Disinfect	Betaisodona tool	Conjunctiva and cornea
n°4	Incise	1.1 mm knife	Cornea
n°5	Irrigate	Sauter cannula	Conjunctiva and cornea
n°6	Hold	Colibri tweezers	Bulbus oculi
n°7	Incise	1.4 mm knife	Cornea
n°8	Irrigate	Irrigation cannula	Anterior chamber
n°9	Inject	Methocel tool	Anterior chamber
n°10	Cut	Wecker scissors	Lens
n°11	Phacoemulsificate	Chopper	Lens
n°12	Hold	Micro spatula	Bulbus oculi
n°13	Aspirate	Syringe	Anterior chamber
n°14	Aspirate	Aspiration cannula	Lens
n°15	Irrigate	Irrigation cannula	Lens
n°16	Implant	IOL tool	Anterior chamber
n°17	Place	Reposition hooklet	Lens

III.2.d. Visualization of surgical processes

Once the formalization and the definition of activities have been performed, a visual representation of the data is required to easily explore them and to illustrate results. However, complex data structures sometimes prevent straightforward visualization. In the case of SPs, we propose the use of index plots (Scherer, 2001), which have already been used for sequence visualization (Brzinsky et al., 2006). The idea of an index plot is to display the sequence by representing an activity as a rectangle of a specific color for each activity, and a width proportional to its duration ($i.e.\ stop(ac_i) - start(ac_i)$). By this mean, SPs can easily be visualized and a qualitative visual comparison can be performed. The following gives an example of one SP.

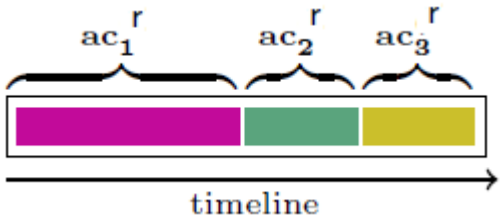


Figure 16 - Example of 3 activities of the right hand of one SP.

Using this technique, the activity recording of the entire surgery can be represented using index-plots (**Figure 17**).

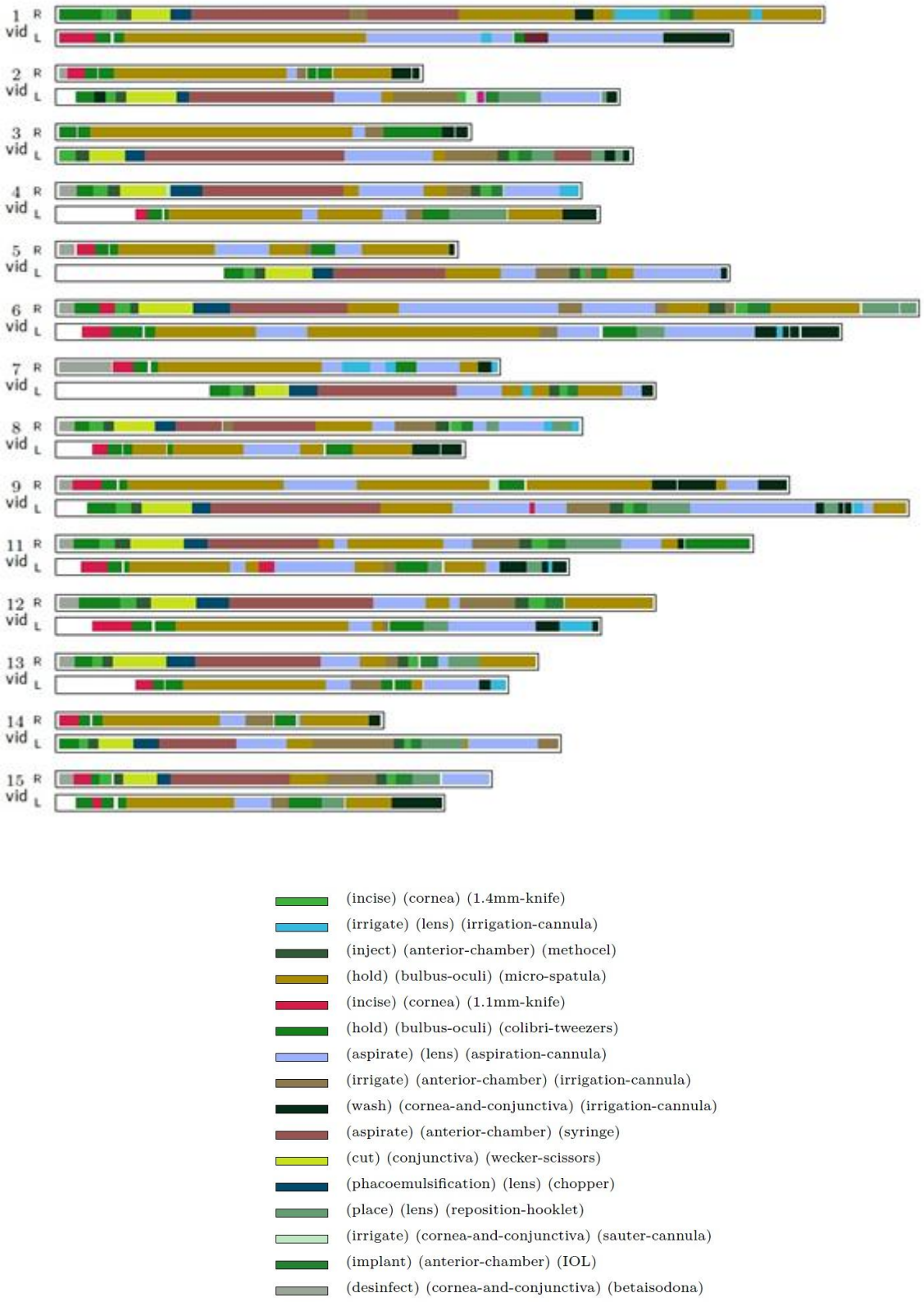


Figure 17 - Index-plot visual representation of 15 videos of cataract surgeries and the colour legend.

III.3. Discussion

III.3.a. Choice of surgical procedures

We choose pituitary and cataract surgeries due to their relative good reproducibility. Both surgeries are common procedures where few particular adverse events can occur and where surgeons always follow identical surgical processes. Even if differences between surgeons can be found, these differences are minimal. It can be differences in term of surgical time or dexterity but it doens't affect the surgical processes, being at a high- or low- granularity level. Moreover, they are also very standardized procedures, already widely studied by both the clinical community and the methodological community. A kind of consensus on the terminology for describing these surgeries has emerged and this standardization is very important for sharing ideas and results between research teams.

The choice of reproducible and standardized procedures is not restricted to these two examples. In neurosurgery, we also collected a set of cervical disc herniation removal procedures by antero-lateral cervicotomy (mean time of microscope use ~ 45min) that could have been included for further analysis. Due to the small number of cases that we collected (<10) for this type of surgery, we decided not to perform experiments on this dataset. Concerning the number of cases for pituitary (16 videos) and cataract (20 videos) surgeries, it is sufficient to perform cross-validation studies but studying more cases is of course always better. It would allow a better evaluation of systems performance, and we could imagine that around 100 videos per surgery could be a solid base for accurately evaluating performances.

III.3.b. Identification of high- and low- level tasks

During the identification of high-level tasks, no formalization was asked for the definition of the each phase. On the contrary of low-level tasks that had a strong formalization, free text was used for defining each surgical phase. For the dataset of pituitary surgery videos, we decided to fuse the initial possible "access to the tumour" and "tumour removal" phases because, for this type of surgical procedure, it is currently hard to distinguish them with image-based algorithms alone. The transition between both is not clearly defined due to the similar tools and microscope zooms used while performing these tasks. Others 5 surgical phases follow the overall workflow of usual pituitary surgery.

For the dataset of cataract surgery videos, three high-level tasks that can be seen as overall surgical phases have been merged with other phases: viscoelastic injection, performed just before capsulorhexis to stabilise the anterior chamber, hydrodissection just before phacoemulsification and wound sealing at the end of the operation, providing complete waterproofness of the anterior chamber. With an image-based analysis, these phases are quite impossible to detect because very small instruments are used and there are no major change in colour, form or texture. That is why we merged them with their consecutive phases. Other common tasks frequently performed by a surgeon during cataract surgery are hydration of the eye and the act of sponging, which generally occur more than ten times during a normal procedure. These tasks also were not detected, as they cannot be considered as real phases that follow a sequence.

References

- ✓ Brzinsky-Fay C, Kohler U, Luniak M. Sequence Analysis with Stata. Stata Journal. 2006; 6(4): 435-60.
- ✓ Ezzat S, Asa SL, Couldwell WT, Barr CE, Dodge WE, Vance ML, et al. The Prevalence of Pituitary Adenomas: a systematic review. Cancer. 2004; 101(3): 613-22.
- ✓ Neumuth T, Strauß G, Meixensberger J, Lemke HU, Burgert O. Acquisition of Process Descriptions from Surgical Interventions. DEXA 2006: Proc Int Conf on Database and Expert Systems Applications; LNCS 4080: 602-11.
- ✓ Neumuth T, Trantakis C, Eckhardt F, Dengl M, Meixensberger J, Burgert O. Supporting the analysis of intervention courses with surgical process models on the example of fourteen microsurgical lumbar discectomies. Comput Assisted Radiol Surg 2007: 436-8.
- ✓ Neumuth T, Jannin P, Strauss G, Meixensberger J, Burgert O. Validation of Knowledge Acquisition for Surgical Process Models. J Am Med Inform Assoc. 2008; 16(1): 72-82.
- ✓ Scherer, S. Early career patterns: A comparison of Great Britain and West Germany. Eur Sociol Rev. 2001; 17(2): 119-44.

Chapter IV. Surgical phases detection *static approach*

IV.1. Introduction

The challenge we aimed at addressing in this Chapter was to study if surgical phases can be automatically recognized from video images only. The first method we studied consisted in classifying each frame of surgical microscope videos independently without taking into account the sequential aspect of the surgical phases. We addressed such static approach as an image classification problem. Image features are first extracted from each frame in order to create image signatures. After a step of feature selection, a supervised classification is performed to assign each frame to a surgical phase. This chapter is organized as follow. We first propose a specific state-of-the-art of SPMs using on-line video-based recording, and give a short overview of the main approaches for low-level feature extraction in image processing. Then, we present the method we proposed and the validation studies we conducted using the pituitary data-set. We finally discuss some important aspects of this method.

IV.2. SPMs using on-line video-based recording

Recently, several studies focused on the recognition of surgical tasks in the OR from videos. Based on **Table 3**, 10 studies used on-line video-based recording. These studies can be distinguished according to the type of videos used: studies using external OR videos and studies using endoscope videos.

Firstly, the use of external OR videos has been tested. Bhatia et al. (2007) analysed overall OR view videos. After identifying 4 phases of a common surgical procedure, relevant image features were extracted and HMMs were trained to detect OR occupancy. Padoy et al. (2008, 2010) also used low-level image features through 3D motion flows combined with hierarchical HMMs to recognize on-line surgical phases. Hu et al. (2006) combined patient vital signs with visual features for detecting OR status. Suzuki et al. (2010) used multichannel visual information and quantified the motion in the OR using video file size for having an idea of the progression phase of the intervention.

Secondly, the use of endoscope videos in MIS has been mainly investigated. The main constraints in MIS range from the lack of 3D vision to the limited feedback. However, studies on the subject have recently shown that the use of videos in this context was relevant. Speidel et al. (2008) focused on surgical assistance for the construction of context-aware systems. Their analysis was based on augmented reality and computer vision techniques. They identified two scenarios within the recognition process: one for recognizing risk situations and one for selecting adequate images for the visualisation system. Lo et al. (2003) used vision and particularly visual cues to segment the phases. They used colour segmentation, shape-from-shading techniques and optical flows for instrument-tracking. These features, combined with other low-level visual cues, were integrated into a Bayesian

framework for classification. Klank et al. (2008) extracted image features for further scene analysis and frame classification. A crossover combination was used for selecting features, while Support Vector Machines (SVMs) were used for the supervised classification process. Also in the context of endoscopic interventions, Blum et al. (2010) automatically segmented the surgery into phases. A Canonical Correlation Analysis was applied based on tool usage to reduce the feature space, and resulting feature vectors were modelled using DTW and HMM. James et al. (2007) used visual features combined with an eye-gaze tracking system following the eyes movements of the surgeon to detect one important phase (i.e., clipping of the cystic duct).

A third family of video-based analysis can also be defined: videos coming from robot in the context surgical skill evaluation (see subsection II.3). With videos coming from the Da Vinci robot, Voros and Hager (2008) used kinematic and visual features to classify tool/tissue interactions in real-time. Similarly, Reiley and Hager (2009) focused on the detection of activities for surgical skill assessment.

IV.3. Low-level image features extraction

A vital requirement for reliable vision systems is the ability to extract relevant spatial and temporal image features from the video. The step of image feature extraction is traditionally included in more complex processes of multimedia data-mining. It allows the input image to be transformed into a reduced representation set of features. Multiple axes have been proposed in computer vision for extracting information in colour images. The work of Marr (1982) first defined the different steps toward the creation of a computer vision system, from low-level features extraction to high-level interpretation of the image. The gap between these two levels has been widely addressed and is called "the semantic gap". Moreover, the increasing number of multimedia in daily life has motivated researches on Content-Based Image Retrieval (CBIR). The goal is not to give an exhaustive list of current methods for extracting features from images, but to introduce the main approaches. In this subsection, we focus on the static low-level feature extraction process. Such image features are extracted using various methods that can be differentiated by two aspects: the type of features (color, texture or form) and the method for the extraction (global or local approach).

Color is one of the primary visual features used to represent and compare visual content. Its study has been an active area of research in image processing. An image can be described as linear combinations of 3 basis colors called primaries. It is a subjective characteristic that is often represented with color histograms. Challenges and problems of image classification using color histograms have been discussed by Gong et al. (1998). For the representation, different color spaces have been proposed. The RGB (Red Green Blue) representation is closely relates to human perception, as it is designed to match the input channel of the eye. It is the first space that has been employed by the community and that showed satisfactory results in image classification problems. The limitation of this representation is that the three components are not independent. Then, the HSV (Hue-Saturation-Value) or HSL (Hue-Saturation-Lightness) representation has been selected for its invariant properties. It's a cylindrical coordinate representation of the RGB space. The chrominance components (Hue-Saturation) are invariant to changes in illumination intensity and shadows. Moreover, the hue is invariant under the orientation of objects with respect to the illumination and camera direction and hence more suited for object retrieval. It is also well adapted for multi-resolution, where data can be analyzed with fine details. Other color spaces, using the same principles that these two main

representations have also been proposed: the CIE LUV (Park et al., 1999), the HVC (Gong et al., 1998), or the YIQ and the YUV used for television. Fusion or mix of these representations may be adapted for correctly representing an image through its color.

Texture is another type of feature that can be extracted from images. The texture is defined as a tactile or visual characteristic of the surface. It generally captures patterns or lack of patterns in the image data. Statistical measures are simply used, e.g. entropy, homogeneity or contrast. The co-occurrence matrix of the Haralick descriptors (Haralick et al., 1973) has been also mostly employed. Another important texture extraction technique is the use of multi-resolution simultaneous autoregressive models-MRSAR (Jianchang and Jain, 1992). Gabor filters (Feichtinger and Strohmer, 1998) have also been used, where statistical features (e.g. average, variance) are extracted from the image after applying the Gabor filter. Other ones use Markov models (Ashlock and Davidson, 1999). The use of wavelets has also gained attention (Daubechies, 1992) due to their local properties, as well as the use of fractals (Varma and Garg, 2007).

Form is the third main feature that can be extracted from images. It is the most important visual feature that can describe an image. However, the description of form is a difficult task, as it requires a preliminary segmentation step that may not be reliable. That's the reason why many current systems don't use this type of features. When used, the form is often characterized using invariant moments. The most employed ones have been the Hu (Hu, 1962) moments and the Zernicke moments (Teague, 1979).

In addition to these three components, the use of image-processing operations that transpose the image into another representation can be used. Once the transformation is performed, features can be extracted on the new representation space allowing the accentuation of specific features. Algorithms like the Discrete Fourier Transform, Fourier-Mellin (Derrode and Ghorbel, 2001), Principal Component Analysis, Discrete Cosine Transform (Ahmed et al., 1974) are possible ways of transforming the image for further extraction of features.

These different types of low-level features can be extracted either using a **global approach** over the entire image or a local approach after a pre-segmentation of the image. A system based on global features only cannot capture the local information (e.g., objects) and provides a rough representation of the content of the image only. On the contrary, the **local approach** alone doesn't preserve the significance and the coherence of the global image. A compromise has to be found between both approaches according to the application. Within local approaches, two methods are usually employed. The first one, which is the easiest one, consists in the division of the image using a grid where features are then computed for each region. The second one consists in the segmentation of the image in order to partition the image into objects and create areas of similar image features. The main issue of this approach remains the choice of the segmentation method which is never perfect and which may impact the feature extraction.

A large number of algorithms have been proposed for extracting low-level image features, but the choice of a particular one for a specific image classification problem remains very difficult. The solution that teams have principally followed is the extraction of many features using different complementary algorithms of colour-, texture- and form- oriented features and then to launch feature

selection algorithms that are able to select the most discriminant ones that are adapted to the problem. That's the solution that we will also follow in this Chapter.

IV.4. Methods

We present here the methodology for classifying video frames using a static approach. After a step of pre-processing, a feature extraction process was first performed for each frame, resulting in image signatures composed of 185 complementary features. Discriminant ones were chosen with a specific feature selection method that combines a *filter* and a *wrapper* approach. Supervised classification was then used to classify the frames. The framework is shown on **Figure 18**. We assessed this process with cross-validation studies. These different steps are described in the next sub-sections.

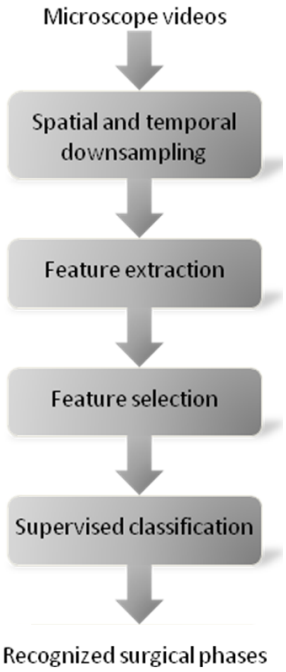


Figure 18 - Workflow of the recognition process.

IV.4.a. Pre-processing

Frames were first sequentially extracted from microscope videos and pre-processed. The video sequences were downsampled to 1 frame every 2s (0.5 Hz). We also spatially downsampled frames by a factor of 8 with a 5-by-5 Gaussian kernel[2].

[2] Internal studies have shown that up to this downsampling rate, there was no impact on the classification process

IV.4.b. Feature extraction

We defined for each video frame a feature vector that represented a signature. According to the overview of low-level image feature perform in subsection IV.3, image signatures were composed of the three main types of information that usually describe an image: colour, texture and form (**Figure 19**). The most common methods of low-level image feature extraction were chosen.

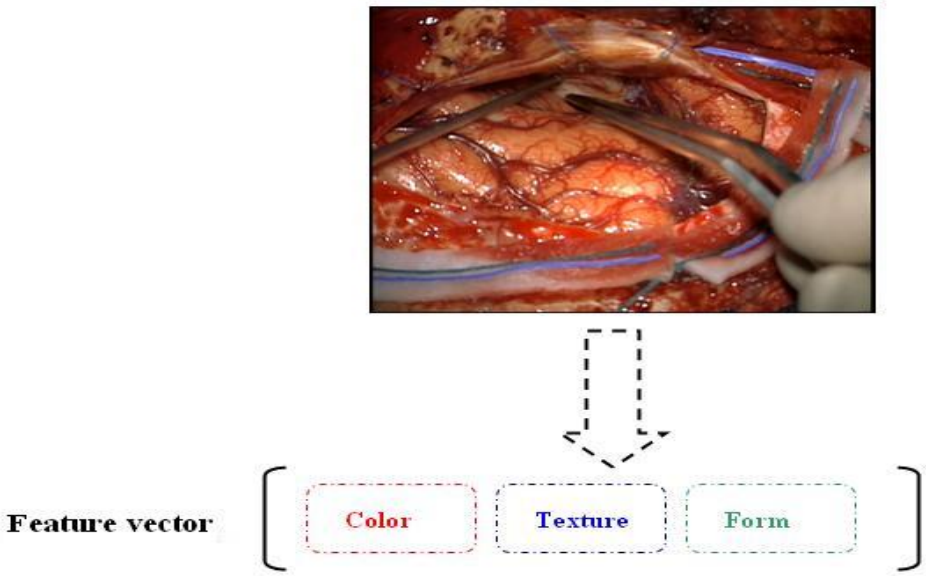


Figure 19 - Feature vector (i.e. image signature) for one frame of the pituitary data-set.

Colour was extracted from two complementary spaces (Smeulders et al., 2000): RGB space (3 x 16 bins) along with Hue (30 bins) and Saturation (32 bins) from HSV space.

Texture was extracted with the co-occurrence matrix along with Haralick descriptors. The co-occurrence matrix C was used to describe the patterns of neighbouring pixels in an image I at a given distance. Mathematically, the matrix C was defined with an image n x m and an offset:

$$C_{\Delta x,\Delta y}(i,\ j) = \sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, if\ I(p,q) = i\ and\ I(p+\Delta x, q+\Delta y) = j\\ 0, otherwise\end{cases} \quad (4)$$

Four matrices were computed for different orientations (horizontal, vertical and two diagonal directions). A kind of rotation invariance was achieved by taking the four matrices into account. Haralick descriptors were then used by computing the contrast, the correlation, the angular second moment, the variance of the sum of squares, the moment of the inverse difference, the sum average, the sum variance, the sum entropy, the difference of variance, the difference of entropy and the maximum correlation coefficient of the co-occurrence matrix.

Form was represented with spatial moments describing the spatial distribution of values. For a greyscale image, the moments $M_{i,j}$ are calculated by:

$$M_{i,j} = \sum_{p=1}^{n}\sum_{q=1}^{m} p^i q^j I(p,q) \quad (5)$$

The 10 first moments were included in the signatures. We then computed the Discrete Cosine Transform (DCT) (Ahmed et al., 1974) coefficients B_{pq} that reflect the compact energy of different frequencies. DCT is calculated by:

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos\frac{\pi(2m+1)p}{2M} \cos\frac{\pi(2n+1)q}{2N}, \quad \begin{matrix} 0 \le p \le M-1 \\ 0 \le q \le N-1 \end{matrix} \tag{6}$$

where $\alpha_p = \begin{cases} 1/\sqrt{M}, & p = 0 \\ \sqrt{2/M}, & 1 \le p \le M-1 \end{cases}$ and $\alpha_q = \begin{cases} 1/\sqrt{N}, & p = 0 \\ \sqrt{2/N}, & 1 \le p \le M-1 \end{cases}$

The B_{pq} coefficients in the upper left corner represent visual information of lower intensities, whereas higher frequency information is gathered in the right lower corner of the block. Most of the energy is located in the low frequency area, which is why we took the 25 features in the upper left corner. After this step, each image signature was finally composed of 185 complementary features.

IV.4.c. Feature selection

It is well established that the use of too many variables in a classification procedure may decrease classification accuracy. Therefore, image signatures have to be reduced to improve classification results but also to decrease computation time. This step has to be performed once only for each learning database, and then selected features are used for each recognition run. Two techniques are mainly used for this task: Principal Component Analysis (PCA) and feature selection. We tested both methods on our data.

PCA (Jolliffe, 1986) is a statistical method used to decrease the data dimension while retaining as much as possible of the variation present in the dataset to process the data faster and more effectively. The set of observations is transformed into a new space with uncorrelated variables, named as the principal components, each of which is a linear combination of the original variables. The first principal component is computed to have the highest variance (i.e. to capture the variability of the data), and succeeding components are also ranked according to their variance. Most of the time, the first few components capture the majority of the observed variation.

Similarly, the main goal of feature selection methods (Saeys et al., 2007) is to remove redundancy information and to keep the essential features for the recognition step that follows. The resulting chosen features are often a good compromise between computation time and recognition performance. In order to decrease the data dimension, and knowing that too many features can decrease the correct classification rate, we performed feature selection studies to find the best combination of features. Typical feature selection methods can be divided into two groups (Duda and Hart, 1973) depending on their evaluation procedure, the filter and the wrapper methods. Algorithms from these groups can be supervised or non-supervised. Filter techniques do the feature selection by looking at the intrinsic properties of the data only. They are therefore independent of the inductive algorithm, which is a disadvantage. In wrapper methods (Kohavi and John, 1997), various subsets of features are generated and evaluated. A classification algorithm is used as the outcome for evaluation. We fused these two feature selection algorithms using the method described in Mak and Kung (2008). In their work, they

argued that both types of selection techniques are complementary to each other. Two algorithms are independently applied to find two feature subsets of identical size. They are then merged by selecting one feature at a time from both subsets, starting with the highest ranking feature. The final feature subset is then a combination of results from one filter and one wrapper. The Recursive Feature Elimination (RFE) SVM (Guyon et al., 2002) was chosen for wrapper methods. The principle is to generate the ranking of features using backward feature elimination. The mutual information (Hamming, 1980) represents the filter methods. A feature is more important if the mutual information between the target and the feature distributions is larger.

IV.4.d. Supervised classification

With the representation of image into descriptor vectors, we are now able to train classification algorithms following the principle of an image classification problem (**Figure 20**). The m images are represented by a set of n features (matrix n x m corresponding to the training database of features), and are all described by a set of m classes (matrix m x p corresponding to the training database of classes). Supervised classification is then used for classifying every query image by assigning the p classes to the query image.

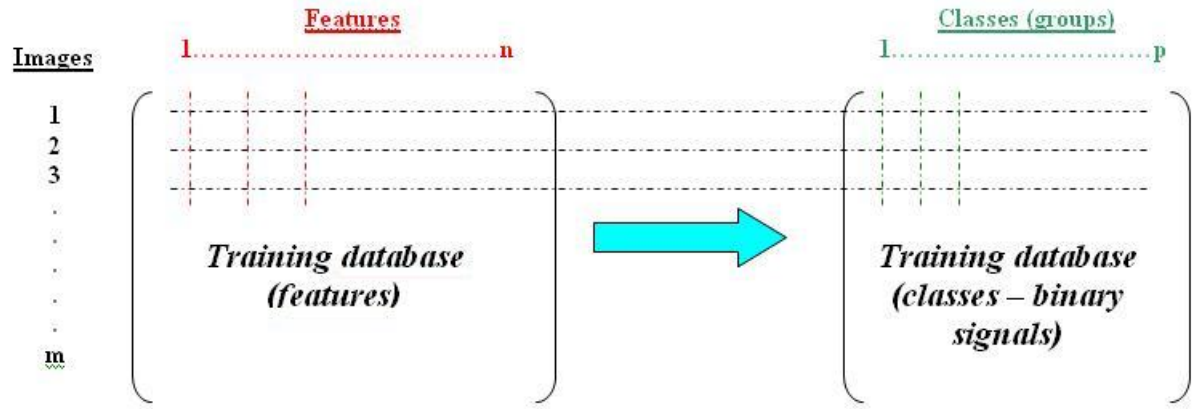


Figure 20 - Training database and the corresponding classes (e.g. phases).

For the recognition process, we tested five classification algorithms: Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Neural Networks (NN), Decision Trees and Linear Discriminant Analysis (LDA). Parameters chosen for these 5 algorithms are presented on **Table 7**. SVMs (Vapnik, 1998) are supervised learning algorithms used for classification and regression. The goal of SVMs is to find the optimal hyperplane that separates the data into two categories. SVMs are often known to produce state-of-the-art results in high dimensional problems. The multiclass SVMs (Crammer and Singer, 2001) extends it into a K-class problem, by constructing K binary linear SVMs. Mathematically, given training data $(x_1 ... (x_{(n}))$ where $x \in \Re^d$ and their labels $(y_1 ... (y_n))$ where $y \in (-1,1)$, the goal is to find the optimal hyperplane $w.x + b = 0$ that separates the data into two categories. The idea is to maximize the margin between the positive and negative examples. The parameter pair $(w;b)$ is finally the solution to the optimization problem:

$$\varphi(w) = \frac{1}{2}\|w\| = \frac{1}{2}(w.w) \tag{7}$$

following constraints:

$$y_i(x_i.w+b)-1\geq 0, \forall i \quad (8)$$

The KNN algorithm (Dasarathy, 1990) is the simplest method for classification. Each point in the space is assigned to class C if it is the most frequent class label among the k-nearest training samples. NNs (Haykin, 2008) are non-linear statistical methods inspired by biological neural networks. They are often used to model complex relationships between inputs and outputs. We used these in a supervised way with a back-propagation neural network. The Decision Tree (Breiman et al., 1984) is a quick classification algorithm where each internal node tests an attribute. It is notably used when data are noised and classes are discrete. Lastly, LDA (Mclachlan, 2004) is based on a Fisher analysis. It is a linear combination of features that best separates two or more classes.

Table 7 - Parameters of the 5 classification algorithms tested for extracting binary cues.

Algorithms	Parameters
SVM	Kernel : Linear
KNN	Distance : Euclidean
Neural Network	Feed-forward back propagation network
	Transfer function : linear
	Training algorithm : Levenberg-Marquardt
Decision tree	Criterion : Gini index
LDA	Discriminant function : Linear

IV.4.e. Validation studies

Two studies were performed in order to assess this first approach toward automatic surgical phase detection. The first study was performed to select the optimal data dimension reduction technique along with the optimal number of image features for image signatures. With the best settings found, we also studied classification algorithms for the recognition of the surgical phases. The feature selection method was used to select the most discriminant features between phases.

Both studies were performed using a random 8-fold cross-validation (Duda et al., 2001). From the initial image database of 16 videos of pituitary surgeries, the dataset was divided into 8 random subsets (i.e. 2 videos and their corresponding frames per subsets), where seven were used for training while the prediction was made on the eight's subset. The procedure was repeated 8 times and accuracies were averaged. In addition, sensitivity and specificity were also computed.

Accuracy was defined by $Acc = \frac{TP+TN}{TP+FP+FN+TN}$ and represented the percentage of correctly classified frames. Specificity was defined by $Spe = \frac{TN}{TN+FP}$ and sensitivity by $Sen = \frac{TP}{TP+FN}$, where FP is False Positive, defined as images belonging to phase i not identified as part of phase i, TP is True Positive, FN is False Negative, defined as images not belonging to phase i identified as part of phase i, and TN is True Negative.

IV.5. Results

Figure 21 shows that PCA was more suitable for the recognition of binary cues when fewer than 30 features were kept for the classification. From this threshold, the hybrid feature selection gave the best accuracy and reached its maximum for a number of 40 features. These results are borne out with the two classical classifiers, the KNN and the SVM. When using PCA and a SVM classifier specifically, the accuracy sharply decreased from 40 features, whereas with the same classifier but the other feature selection method the accuracy was almost unchanged and stayed at a high recognition rate. When using a KNN, neither data dimension reduction method had any major impact on accuracies, which stagnated from 40 features.

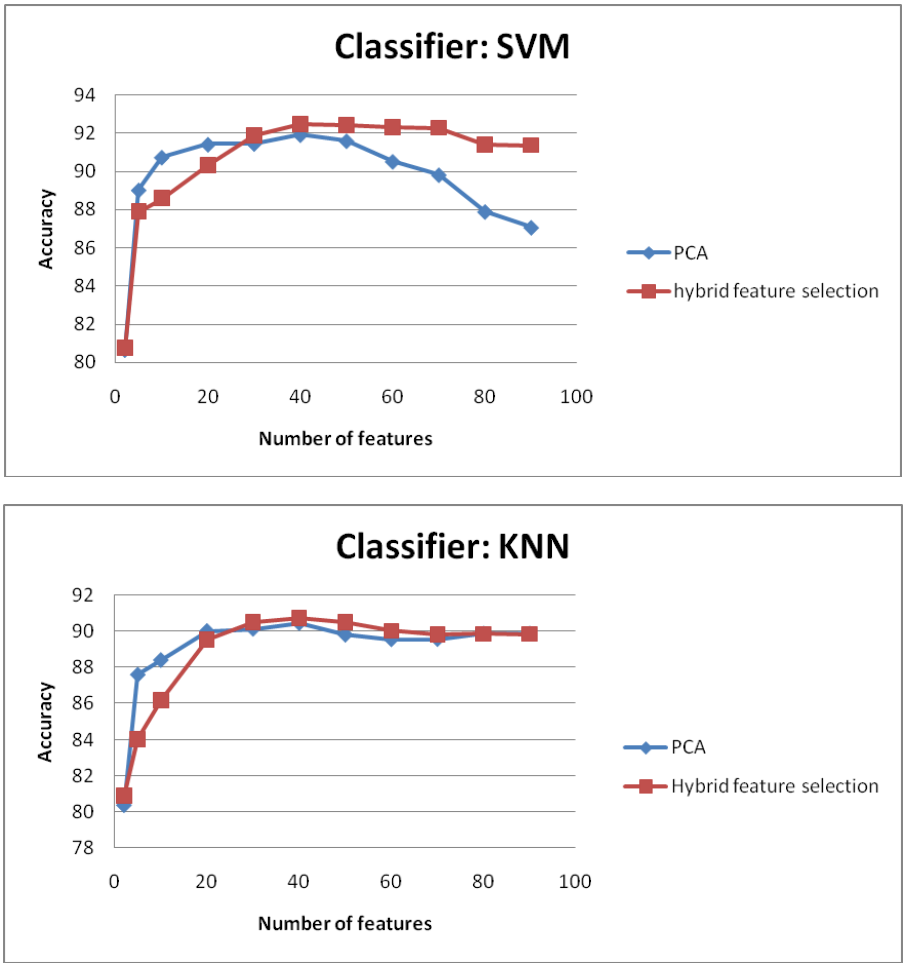


Figure 21 - Correct classification rate (accuracy) according to the number of features kept for two classifiers (SVM and KNN), with two different data dimension reduction methods (PCA and hybrid feature selection).

When applying a PCA transformation on a set of data, it's often useful to have an idea of the energy (i.e. variance) when using the new representation space for each number of features (**Figure 22**). For instance, using our data, the 40 principal components represented 91.5% of the total energy.

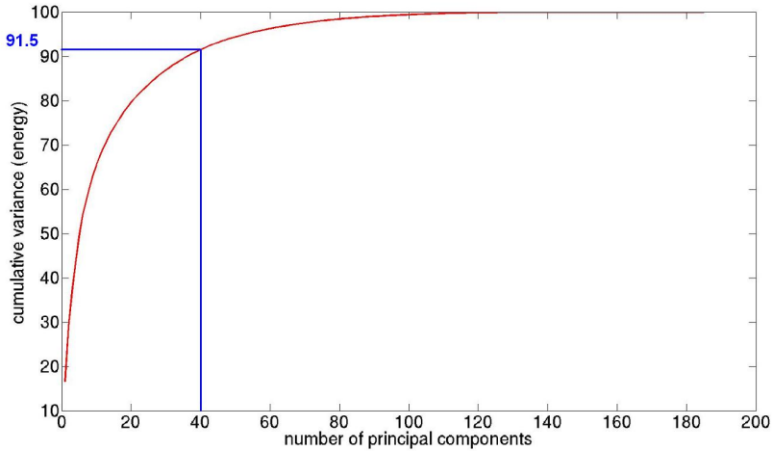


Figure 22 - Cumulative variance of the PCA.

With 40 features taken from the hybrid feature selection method (according to results of **Figure 21**), the other three classifiers were tested in **Table 8** in order to find the best classifier. The SVM classifier gave the best results (91.5%), with good detection accuracy for the LDA and KNN classifiers. On the other hand, the decision tree and NN gave the worst results. We also can see from **Table 8** that specificity was always greater than sensitivity for all algorithms. Maximum specificity was obtained when using SVM (95.2), whereas LDA seemed to be robust when high sensitivity was needed (72.3%). The decision tree showed its limits (specificity: 58.8%).

Table 8 - Correct classification rate (accuracy), sensitivity and specificity of classification algorithms. Image signatures are composed of the 40 first principal components.

Algorithms	Accuracy	Sensitivity	Specificity
Multiclass SVMs	82.2%	78.7%	98.1%
KNN	74.7%	66.0%	95.4%
Neural Network	71.3%	65.1%	92.8%
Decision tree	66.2%	52.3%	94.0%
LDA	81.5%	77.0%	97.6%

IV.6. Discussion

Our global workflow, including image database labelling, features extraction, feature selection, and supervised classification, makes possible the recognition of surgical phases. After experiments, we finally kept 40 features from the hybrid feature selection method and the multiclass SVMs as supervised classifier for a global accuracy of 82%.

IV.6.a. Data dimension reduction

Two methods were tested for reducing the initial image signature dimension: a PCA and a hybrid method combining a wrapper with a filter approach. Intuitively, wrapper approaches seem more advantageous, since the image features are selected by optimising the discriminative power of the

underlying classifier. The main drawback of such methods is the large amount of computation needed to cover the entire search space. In the filter approach, features are selected with no regard to what classifier will be adopted, by evaluating the individual predictive power of each feature. There are thus some limitations, given the fact that the evaluation is performed without any knowledge of class labels. The algorithm does not take into account the power of discrimination of feature combinations, affecting classification accuracy.

The value of combining both methods is to leverage advantages from both to achieve the optimum strategy. We also performed this selection strategy in order not to bias the classification studies. As many machine learning algorithms were tested for evaluating the binary cues extraction, the choice of taking only a wrapper method with a specific classifier would have affected the results by clearly emphasising this classifier in the classification accuracy computation. Moreover, **Table 8** shows that SVMs outperform other classifiers, a result that could be explained by the selection of SVM-RFE as the wrapper method. Nevertheless, internal studies have shown that whatever the wrapper method chosen, it had no impact on the superiority of SVM performances compared to others. **Figure 21** also shows that the hybrid feature selection method outperforms PCA for data dimension reduction. The main limitation of PCA as the filter method is that it makes no use of class labels. The hybrid feature selection method has thus been used for the rest of this first study.

IV.6.b. Explanation of classification errors

From the pituitary surgery data-set, we decided to fuse the initial possible phases "access to the tumour" and "tumour removal" since it's currently hard to distinguish them with image features only, for this type of surgical procedure. The transition between both is not clearly defined due to similar tools and same microscope zoom values used while performing these tasks.

The correct classification rate includes the results of the cross-validation study for the six phases. From these results, we noticed frequent confusions mainly between phase n°3 and n°4, and also between n°1 and n°5. These errors are explained by the very close image features of these phases. Same microscope zoom values, along with similar colours and same surgical instruments make the recognition task very difficult. One solution of this issue would be to integrate one other signal: the surgery time. This information would for instance permit to correctly recognize an image originally identified as part of phase n°5 or part of phase n°1. On the other hand it would still be hard to separate consecutive phases.

IV.6.c. Classification algorithms

In the binary cues extraction study, SVMs and LDA gave the best correct classification rates. The high accuracy of SVMs, associated with small standard deviation, indicates that SVMs are very robust for microscope image classification. Additionally, a validation study using a Gaussian kernel has been performed, showing no differences compared to a linear kernel. LDA is, like many others, optimal when features have a normal distribution. Results have also shown that this classifier was well suited to this setting. On the other hand, the decision tree, NN and KNN gave worse results. Our dataset was probably too variable (in colour and texture) and not discriminant enough to train accurate models with decision trees and KNN. NNs were quite surprising in their ability to improve their performances when the amount of data increases. Non-linear algorithms are generally more suitable for complex systems, which is not the case here. On the other hand, linear algorithms are more straightforward and easy to use, which is why they seem to be more adaptable for our system.

The correct classification rates for SVMs, KNN and NN are almost constant up to 185 features, whereas the accuracy of the decision tree and especially LDA decreases. This is due to the high dimension of inputs, which usually decreases the results of classifiers. It is also why we only kept 40 features for image signatures. If data reduction had not been performed, we would only have obtained an accuracy of 78% (best result obtained with KNN), which demonstrates the usefulness of this step in our workflow.

According to **Table 8**, most of the difference between classifiers is made by sensitivities, which are lower than specificities. High specificity is due to the absence of false positives, whereas low sensitivity is due to a high false negative rate. Thus, the challenge in the future will be to decrease FN rates. When unexpected events occur, such as bleeding or a sudden microscope move, specificity decreases slightly and thus affects overall accuracy. As such situations are unpredictable, they sharply limit classification from static images only.

IV.6.d. From static approach to dynamic approach

With large images database of the same surgery, we are now able to recognize surgical phases of every query video frame, by computing every frame signature and then launching machine learning techniques. We have validated our methodology with the pituitary surgery data-set, but it can easily be extended to other type of interventions. This type of recognition process is a first step toward the construction of CAS systems based on automatic signals extraction. Unfortunately, the major limitation of this approach remains the lack of sequential information within the recognition process. Confusions between distant phases can occur. For instance, a frame belonging to the first phase can be classified as being part of the last phase, avoiding any application to be considered. That's the reason why this first approach, while being important for computer vision experiments, has to be improved. In order to model the entire surgical workflow by taking into account the sequential aspect of the surgical phases, the next chapter will therefore present the use of time-series analysis for incorporating the information of time within the recognition process.

References

- ✓ Ahmed N, Natarajan T, Rao KR. Discrete Cosine Transform. IEEE Trans Computers. 1974; 90-3.
- ✓ Ashlock D, Davidson J. Texture Synthesis with Tandem Genetic Algorithms Using Nonparametric Partially Ordered Markov Models. Evolutionary Computation, CEC. 1999: 157-63.
- ✓ Bhatia B, Oates T, Xiao Y, Hu P. Real-time identification of operating room state from video. AAAI. 2007; 1761-6.
- ✓ Blum T, Padoy N, Feussner H, Navab N. Workflow mining for visualization and analysis of surgeries. Int J Comput Assisted Radiol Surg. 2008; 3(5): 379-86.
- ✓ Bouarfa L, Jonker PP, Dankelman J. Discovery of high-level tasks in the operating room. J Biomed Inform. 2011; 44(3): 455-62.
- ✓ Boucher A, Thi-Lan L. Comment extraire la sémantique d'une image ? 3rd Int Conf: Science of Electronic, Technologies of Information and Telecommunications. 2005.
- ✓ Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression trees. Wadsworth & Brooks/Cole Advanced Books & Software. 1984.
- ✓ Crammer K and Singer Y. On the Algorithmic Implementation of Multi-class SVMs. JMLR. 2001; 2: 265-92;
- ✓ Dasarathy BV. Nearest Neighbor (NN) Norms: NN Pattern Classification techniques. IEEE Computer Society Press.1990.
- ✓ Daubechies I. Ten Lectures on Wavelets. SIAM. 1992.
- ✓ Derrode S, Ghorbel F. Robust and efficient Fourier-Mellin transform approximations for gray-level image reconstruction and complete invariant description. Journal Computer Vision and Image Understanding. 2001; 83(1).
- ✓ Duda RO and Hart PE. Pattern classification and scene analysis. John Wiley & Sons. 1973.
- ✓ Duda R, Hart PE, Stork D. Pattern Classification. John Wiley & Sons. 2001.
- ✓ Feichtinger HG, Strohmer T. Gabor Analysis and Algorithms: Theory and Applications. Birkhäuser. 1998.
- ✓ Gong Y, Proietti G, Faloutsos C. Image indexing and retrieval based on human perceptual color clustering. IEEC: Proc. of Int Conf on Computer Vision and Pattern Recognition. 1998.
- ✓ Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machine. Machine Learning. 2002; 46: 389-422.
- ✓ Haralick, RM., Shanmugam, K., Dinstein, I. Textural features for image classification. IEEE Trans Systems Man Cybernetics. 1973; 3(6): 621-61.
- ✓ Haykin S. Neural Networks and Learning Machines – 3rd edition. Hardcover. 2008.
- ✓ Hu MK. Visual pattern recognition by moment invariants. IRE Trans Information Theory. 1962; 8(2): 179-87.
- ✓ Hu P, Ho D, MacKenzie CF, Hu H, Martz D, Jacobs J, Voit R, Xiao Y. Advanced Visualization platform for surgical operating room coordination. Distributed video board system. Surg innovation. 2006; 13(2): 129-35.
- ✓ Jianchang M and Jain AK. Texture classification and segmentation using multiresolution simultaneous autoregressive models. Pattern recognition. 1992; 25(2): 173-88.
- ✓ Jolliffe T. Principal component analysis. Springer. 1986.
- ✓ Klank U, Padoy N, Feussner H, Navab N. Automatic feature generation in endoscopic images. Int J Comput Assisted Radiol Surg. 2008; 3(3,4): 331-9.

- ✓ Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997; 97(1,2): 273-324.
- ✓ Lo B, Darzi A, Yang G. Episode Classification for the Analysis of Tissue-Instrument Interaction with Multiple Visual Cues. Int Conf MICCAI. 2003.
- ✓ Mak MW and Kung SY. Fusion of feature selection methods for pairwise scoring SVM. Neurocomputing. 2008; 71: 3104-13
- ✓ Marr D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Freeman. 1982.
- ✓ McLachlan GJ. Discriminant analysis and statistical pattern recognition. Wiley Interscience.2004.
- ✓ Padoy N, Blum T, Feuner H, Berger MO, Navab N. On-line recognition of surgical activity for monitoring in the operating room. Proc Conf on Innovative Applications of Artificial Intelligence. 2008.
- ✓ Padoy N, Blum T, Ahmadi SA, Feussner H, Berger MO, Navab N. Statistical modeling and recognition of surgical workflow. Med Image Anal. 2010; 16(3): 632-41.
- ✓ Park D, Park J, Han JH. Image indexing using color histogram in the CIELUV color space. Proc Japan-Korean joint workshop on computer vision. 1999; 126-32.
- ✓ Reiley C.E and Hager G.D. Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. M2CAI workshop. MICCAI. 2009.
- ✓ Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23(19): 2507-17.
- ✓ Smeulders AW, Worrin M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Trans Pattern Analysis Mach Learning Intell. 2000; 22(12): 1349-80.
- ✓ Speidel S, Benzko J, Krappe S, Sudra G, Azad P, Müller-Stich BP, Gutt C, Dillmann R. Automatic classification of minimally invasive instruments based on endoscopic image sequences. SPIE medical imaging – Visualization, Image-Guided Procedures and Modeling. 2009; 7261: 72610A.
- ✓ Suzuki T, Sakurai Y, Yoshimitsu K, Nambu K, Muragaki Y, Iseki H. Intraoperative multichannel audio-visual information recording and automatic surgical phase and incident detection. Int Conf IEEE EMBS. 2010; 1190-3.
- ✓ Teague MR. Image analysis via the general theory of moments. J Optical Society America. 1979; 70(8): 920-30.
- ✓ Vapnik V. Statistical learning theory. Wiley-Interscience. 1998.
- ✓ Varma M, Garg R. Locally Invariant Fractal Features for Statistical Texture Classification. Conf on Computer Vision. 2007; 1-8.
- ✓ Voros S and Hager GD. Towards “real-time” tool-tissue interaction detection in robotically assisted laparoscopy. Biomed Robotics and Biomechatronics. 2008. 562-7.

Chapter V. Surgical steps detection *dynamical approach*

In the previous chapter, we studied an approach for the recognition of surgical phases using static image features only. Performance evaluations have shown promising results. In this chapter, we present the methods we studied to improve the recognition performance. The extension included three aspects. First we integrated in the framework temporal aspects, and particularly sequential aspects. Second, we studied the addition of spatial local features. Third, we studied the addition of temporal features. Since a sequence of surgical phases can be seen as particular cases of a time series, this chapter will start by an overview of current time series analysis methods. Then, we present an overview of methods for local spatial features identification, object detection and recognition, and temporal features. Then we present the approach we implemented, extending the previous framework with additional computer vision techniques, image processing methods, and time-series algorithms. This new framework was validated on the pituitary data-set as well as on the cataract data-set. Discussions on limitations and perspectives of this framework will conclude this chapter.

V.1. Time-series modelling

Time-series are defined in statistics as sequence of data-points measurements along a time period. It is for instance the air temperature computed every minute during a day (=1440 data-points), or the CAC40 values extracted every hour over an entire month (=720 data-points). These both examples are evenly spaced time-series as every interval between data-points is identical. The opposite of evenly spaced time-series are called unevenly spaced. The difference between time-series and all others sequences is that they has a temporal ordering and they represent a stochastic process. The main methods for investigating time-series are presented in the next subsections: the Dynamic Bayesian Networks, including HMMs, Maximum Entropy Markov Models (MEMMs) and Kalman filter models, the Gaussian Mixture Models (GMM), Conditional Random Fields (CRFs), and Dynamic Time Warping (DTW) algorithms. All techniques presented here are dynamic systems, i.e. systems that vary in time. These techniques allow the analysis and the modelling of time-series, but cannot be considered as predictive tools. For being use as a predictive tool, in the context of time-series modelling, we can for instance cite the ARIVA method (Box and Jenkins, 1970) that has been the most current method employed.

V.1.a. Dynamic Bayesian networks

Bayesian Networks

BNs are defined as directed graphical models that represent dependencies between a set of random variables in a probabilistic model. Each node of the model represents a random variable, and the lack of arcs represents conditional independence assumptions. It encodes the local Markov assumption: a variable is conditionally independent of its non-descendant. BNs have recently proven to be of great interest for various applications, and can be extended in the temporal domain using DBNs. For instance, a BN does not work for analyzing a system that changes over time, that's why DBNs were created. DBNs are also directed graphical models but that represent a sequence of variables, i.e. a stochastic process. DBNs are trainable, encode causality in a natural way, and algorithms exist for learning the structure of the networks and doing predictive inference. Particular examples of DBNs are HMMs, MEMMs or Kalman model filters.

Hidden Markov Models

HMMs (Rabiner, 1989) are statistical models used for modelling non-stationary vector times-series. HMMs are graphical models that have a simple definition of independence: Two sets of nodes A and B are conditionally independent given a third set, C, if all paths between the nodes in A and B are separated by a node in C. An HMM is formally defined by a five-tuple (S,O,Π,A,B), where $S=(s_1...(s_N))$ is a finite set of N states, $O=(o_1...(o_M))$ is a set of M symbols in a vocabulary, $\Pi=(\pi(i))$ are the initial state probabilities, $A=(a(ij))$ the state transition probabilities and $B=(b_i(o(k)))$ the output probabilities. Given the observations and the HMM structure, the Viterbi algorithm (Viterbi, 1969) finds the most likely sequence of states. In other words, two probability distributions are studied: $P(s|s')$ and $P(o|s)$. Compared to the DBN, an HMM represents the state of the world using a single discrete random variable, whereas a DBN represents the state of the world using a set of random variables. Hence, a DBN shows how variables affect each other over time, whereas a HMM shows how the state of the system evolves over time. DBNs can learn dependencies between variables that were assumed independent in HMMs, and provide a unified probability model as opposed to having one model per activity as in discriminative HMMs. There are two main drawbacks in using HMMs for time-series modelling. The first one appears when observations are represented with a rich and complex vocabulary that creates overlap between features and therefore decrease the accuracy of the recognition. When the set of possible observations is not entirely known or fully controlled, a parameterization of the features could be useful. The second problem is that HMMs use a generative joint model in order to solve a conditional problem, which is not optimized. The MEMMs, describe in the next subsection, have been created to handle these two drawbacks.

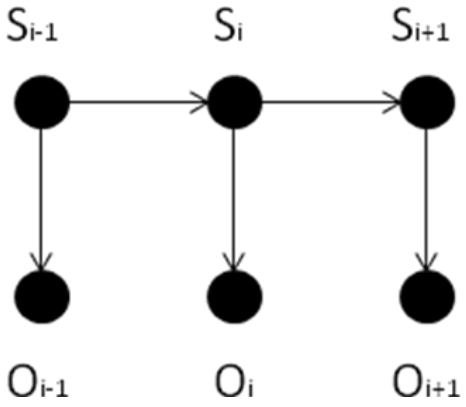


Figure 23 - Structure of a simple HMM.

Maximum Entropy Markov Models

An extension of HMMs is the MEMMs (McCallum, 2001), which do not assume the independency between features and allow observations to be represented as arbitrary overlapping features (**Figure 24**). This model represents the probability of reaching a state given an observation and the previous state, i.e. $P(s|s',o) = P_{s'}(s|o)$. The conditional probabilities that are used in the models are specified by exponential models based on arbitrary observation features. One other advantage of this method is the training, which is very efficient compared to other time-series analysis methods.

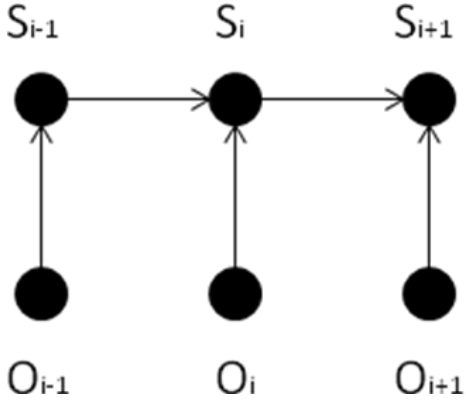


Figure 24 - Structure of a simple MEMM.

Kalman filter models

Kalman filter models (Kalman, 1960), as the HMM, are particular examples of DBNs also used for modelling a discrete time process. They are linear dynamic systems using recursive estimation. For estimating the internal state of a system, the principle is to produce estimates of the true values of noisy measurements by computing weighted averages of the predicted values. This estimation process can be decomposed into two main phases: the prediction and the update that have both dedicated

equations. At each time step, the estimated state from the previous time and the current measurement are provided by minimizing the mean of the squared error. Hence, it support estimations of past, present and future states. Kalman filters present the advantage to support real-time on-going.

V.1.b. Conditional Random Field

CRFs are undirected probabilistic graphical models (Lafferty et al., 2001), in opposition to DBNs that are directed. CRFs have a single exponential model for the joint probability of the entire sequence of labels given the observation sequence. Models define a log-linear distribution of sequences given the observation sequence. One advantage of CRFs compared to MEMMs for example is that they do not suffer from the label bias problem, defined when states with low-entropy transition distribution ignore their observations.

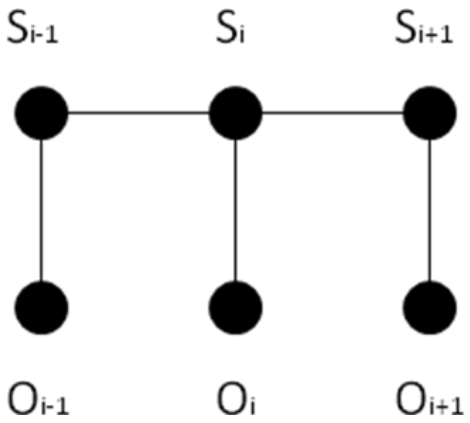


Figure 25 - Structure of a simple CRF.

V.1.c. Dynamic Time Warping

The Dynamic Time Warping (DTW) algorithm (Keogh and Pazzani, 1998) is a method to classify the image sequences in a supervised manner. DTW is a well-known algorithm used in many areas (e.g. handwriting and online signature matching, gesture recognition, data mining, time series clustering and signal processing). The aim of DTW is to compare two sequences $X := (x_1, x_2, ..., x_N)$ of length N and $Y := (y_1, y_2, ..., y_M)$ of length M. These sequences may be discrete signals (time-series) or, more generally, feature sequences sampled at equidistant points in time. To compare two different features, one needs a local cost measure, sometimes referred to as local distance measure, which is defined as a function. Frequently, it is simply defined by the Euclidean distance. In other words, the DTW algorithm finds an optimal match between two sequences of feature vectors, which allows for stretched and compressed sections of the sequence (**Figure 26**).

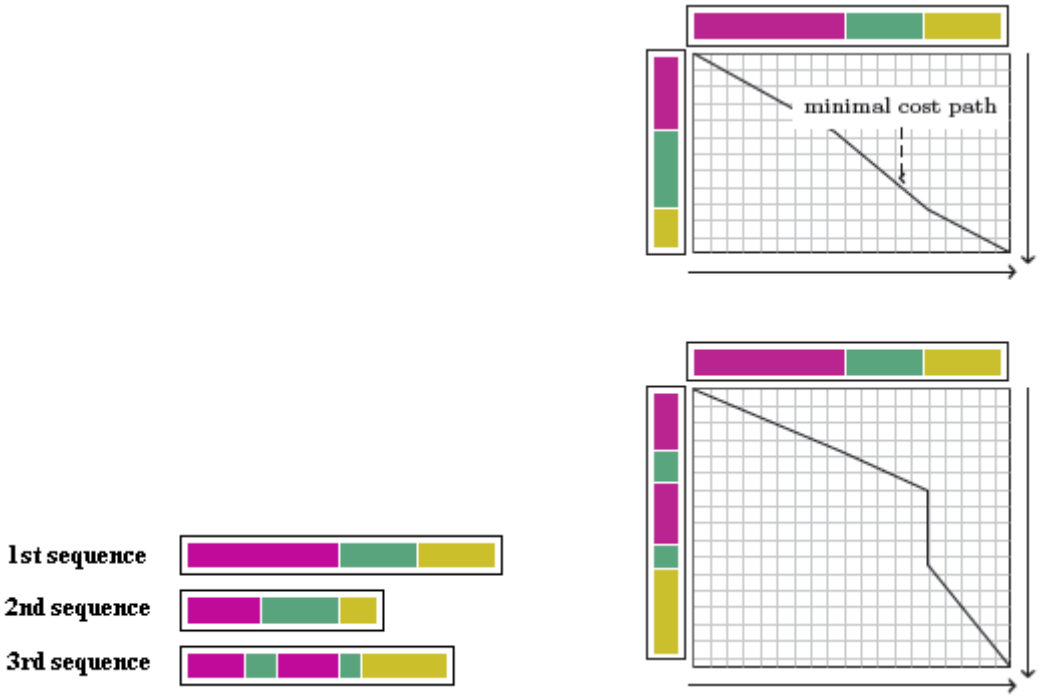


Figure 26 - Minimum cost path for two examples.

Additionally, global constraints (also called windowing functions) can be added to the conventional algorithm in order to constrain the indices of the warping path. With this method, the path is not allowed to fall within the constraints window. The two major constraints in DTW are the Sakoe-Chiba band and the Itakura parallelogram (**Figure 27**).

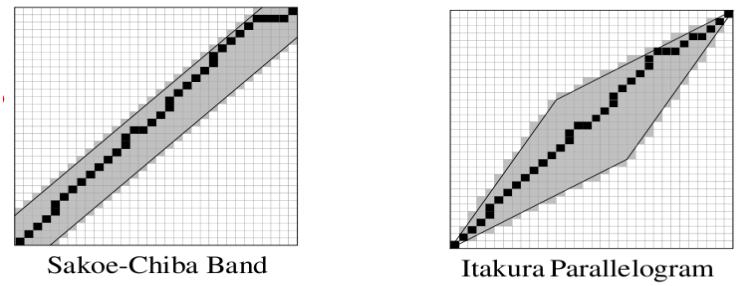


Figure 27 - Two global constraints for the DTW algorithm[3].

V.1.d. Examples of time-series applications

Time-series modelling have been used in different domains that are presented here. For modelling human activities and behaviours, Xiang and Gong (2006) defined activities as a set of discrete events and the modelling is done by reasoning on the temporal correlation between these different events. For

[3] http://izbicki.me

integrating the temporal aspect, authors use derived version of HMMs. In Cuntoor et al. (2008), authors stated that a human activity can be decomposed into a sequence of events having a natural physic interpretation. This decomposition can be performed using a semantic or a statistic approach. The second possibility has been kept, where the modelling is based on the recognition of events from observed data. HMMs were trained for modelling activities. Bhuyan et al. (2006) presented a method for gesture recognition based on the DTW algorithm. Sequences that are used as input for the DTW algorithm are object trajectories representing movements. A sequence of reference was constructed and represented one or multiple gestures. The DTW algorithm was then applied to unknown sequences in order to be timely wrapped to the reference one and recognize the dedicated gestures by transposition.

V.2. Local spatial analysis

The description of the image using low-level image features, as performed in Chapter IV., is intuitively not sufficient, and it is necessary to evolve towards semantic representation of images. Between the extraction of low-level image features and the introduction of semantic, there is a major step. From an algorithmic point-of-view, this step is the transition between basic image processing techniques and more complex computer vision algorithms. From a description point-of-view, this step is the extraction of image features an intermediate-level. While image processing techniques focus on the representation of the image, such as compression or enhancement, and computer vision techniques mainly focus on image understanding, the intermediate-level image features focus on shape-based analysis, edges detection or selection of ROIs. In this section, we present some well known methods for such a purpose. We will therefore investigate the use of edge detection for the creation of masks, morphological operations as pre-processing tools, but also the use of connected component detection for the creation of ROIs. These aspects of intermediate-level image analysis will be used later in the framework to evolve into a more precise description of microscope video frames.

V.2.a. Edge detection

Edge detection is an important part of image processing techniques that convert, in the case of 2D images, an image in a set of curves. The detection of edges is based on the detection of sharp changes in image brightness. Many operators have been proposed for detecting edges, most of them based on gradient operations. All the gradient-based algorithms have kernel operators that calculate the strength of the slope in directions which are orthogonal to each other commonly vertical and horizontal. Here are the most common one:

- ✓ The *Laplacian filter* (Shubin et al., 2001). The locations of the edges can be detected by the zero-crossings of the second-order difference of the image. It is often applied to an image that has first been smoothed with Gaussian filter in order to reduce its sensitivity to noise.

- ✓ The *Canny edge detector* (Canny, 1986). This technique finds the function that optimizes a given functional by the sum of 4 exponential terms. In practice, it can be approximated by the first derivative of a Gaussian.

- ✓ The *Prewitt edge detector* (Senthilkumaran and Rajesh, 2008). It is defined as a discrete differentiation operator that computes an approximation of the gradient.

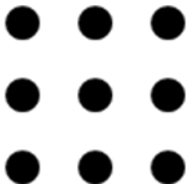
- ✓ The *Roberts edge detector* (Roberts, 1965). The Roberts operator performs a simple, quick to compute, 2-D spatial gradient measurement on an image. Pixel values at each point in the output represent the estimated absolute magnitude of the spatial gradient of the input image at that point.

- ✓ The *Sobel edge detector*. It is very similar to the Roberts edge operator, as it performs a 2D spatial gradient measurement on an image to emphasize regions of high spatial frequency.

V.2.b. Morphological operations

The development of morphological operations has been inspired by image processing issues, where it finds its main application domain. In particular, dedicated tools provide filter, segmentation, and image quantification tools. The idea of mathematical morphology is to study a set of pixels using another set of pixels called the structuring element. In image processing, the first set corresponds to a binary image, and the structuring element to a ROI of size 3x3, 5x5 or 7x7, for instance. The structuring element is located by its centre and is characterized by its shape, its size and its origin. In the case of image processing, the structuring element parameters are defined as:

Size: 3x3, 5x5 or 7x7.
Shape: square
Origin: Middle of the square
Space: discrete



At each position of the structuring element, a response is obtained function of its interaction with the initial set that permits to build the output set. Many operators have been proposed, such as the erosion and the dilatation. Let's denote B the structuring element, which is defined according to its center, E the image space, X the input image. In both techniques, B is moved in order to browse all positions of E.

For the erosion, and for each position of E, we study if B is completely included into E. Mathematically, the positive responses are defined by:

$$E(X,B) = (x \in E, B_x \subset X) \tag{9}$$

For the dilatation, we study if B intersects X. Mathematically, the positive responses are defined by:

$$D(X,B) = (x \in E, B_x \cap X \neq 0) \tag{10}$$

During erosion, the qualitative properties are: the size of objects decreases, small objects and small details disappear, and an object with concavities or holes can be divided into multiples objects. For the dilatation, the properties are: size of objects increases, concavities and holes can be filled in, neighbouring objects can connect with each other. Here are examples of both techniques.

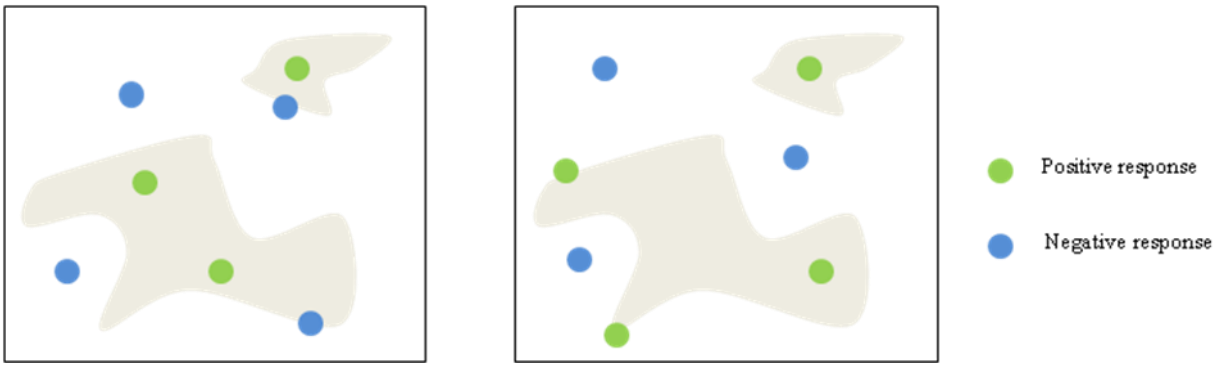


Figure 28 - Examples of erosion (left) and dilation (right) principle (circles=structuring elements).

V.2.c. Connected component detection

This technique, also called in the pattern recognition community *blob extraction* or *region extraction*, uses a sequential algorithm based on a heuristic for labelling connected component elements. In the specific case of image processing, this method allows identifying regions within the image, but is not a segmentation process.

For instance, in the simple case of a binary image and of a 4-neighbor metric (**Figure 29**), two passes are mandatory:

- First pass: the image pixels are scanned from left to right and from top to bottom. For every query pixel of value 1, the left and top pixels are tested:
  - If 2 of the neighbours are 0: assign a new label to the query pixel
  - If only one of the two neighbours is 0: assign the neighbour's label to the query pixel
  - If 2 of the neighbours are 1: assign the left neighbour's label to the query pixel.
- Second pass: For each pixel the smallest label of his neighbour is assign. This step is performed until no more assignment is possible. The different labels found are the connected component of our study.

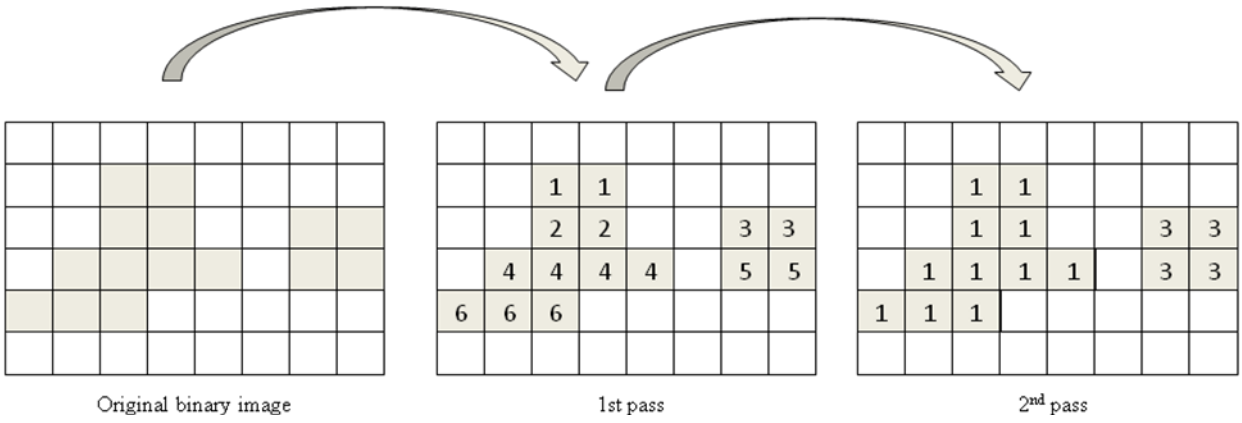


Figure 29 - Examples of a connected component analysis in the case of a binary image and of a 4-neighbour metric.

V.3. Object detection and recognition

One step further for the understanding of images content is object detection and recognition. This is one of the primary and challenging aspects of computer vision. Object recognition can be difficult when the object appears with different scales, orientations, colours, rotations or when it appears partially hidden. Two main approaches have been used for detecting and recognizing objects in images. The first is the *appearance-based approach*, which uses example images of the object (templates) in order to recognize it into new images. From the various methods that have been proposed last few years, we decided to present the two most popular ones: the Haar classifier method and the template matching. The second approach is the *feature-based approach*, which first extracts specific key-points such as corners or edges from images before recognizing objects using supervised classification algorithms. In this category, we will present and use the bag-of-word approach, which has already shown state-of-the-art results in object recognition.

V.3.a. Haar classifier

The Haar classifier (Viola and Jones, 2001), which is a well-know method used by the pattern recognition community, was originally developed for human faces detection, but the method can be used for real-time detection of any types of objects. The Haar classifier is a supervised classifier, which uses the main principles of the Viola-Jones detector. The idea is to create a rejection cascade of nodes (**Figure 30**) where each node is a multitree AdaBoost classifier designed to have high detection and low rejection rates.

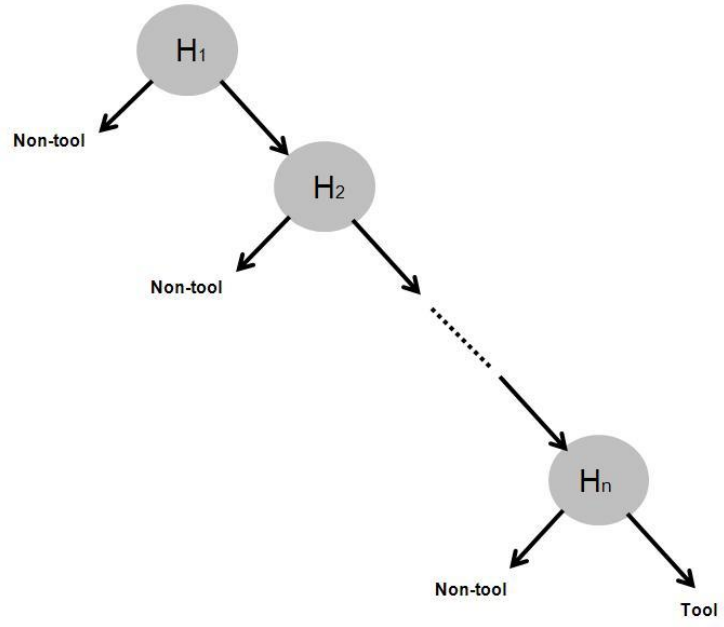


Figure 30 - Rejection cascade used in the Viola-Jones classifier: each node represents a weak classifier tuned to rarely miss a true object while rejecting a possibly small fraction of non-object.

In particular, the use of this method is based on two main stages: the first one consisting in learning the description of the object (training stage) from a database of examples, and the second one consisting in the classification where the object will be recognized in new images.

✓ **Training stage**: The Viola-Jones method is based on the use of boosting to select the best features. Boosting consists of constructing a strong classifier from combinations of weighted weak classifiers. Weak classifiers are Haar-like features. **Figure 31** shows examples of Haar features. This method is based on comparing the sum of intensities in adjacent regions inside a detection window. A black region of the window means "add this area", whereas a light one means "subtracts that area". The method of boosting is the one named AdaBoost proposed by Freund et al. (1995), which combines weighted weak learners to generate a strong learner. Strong learners are then arranged in a classifier cascade tree in order of complexity. The cascade classifier is therefore composed of stages each containing a strong learner. The training needs to be realized over a large number of positive images (i.e. with the object) and negative images (i.e. without the object).

✓ **Detection stage**: the cascade classifier is applied to each window of the query image. Haar-like features used by the current level are computed, and then the classifier response. If the response is positive, the next level is considered and the same computations are performed. If the response is negative, the selected window doesn't contain the object and the next window is treated. The window is declared positive if all levels of the cascade classifier have positive responses, making it through the entire cascade. This method enables a high number of sub-windows to be very rapidly rejected. As a result, the false positive rate and the detection rate are the product of each rate of the different stages.

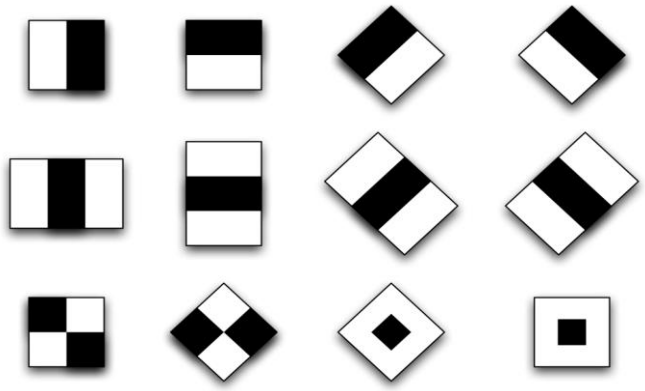


Figure 31 - Examples of Haar-like features[4]

V.3.b. Template matching

The template matching algorithm (Brunelli, 2009), is an algorithm to search areas of an image that match to a template image. To identify the area where the matching is strong, the template image has to be compared with the source image by sliding it. A metric has to be calculated to compare each match at each location. This metric comparing intensities value over a window can be a Sum of Square Difference (SSD), a correlation or other intensity-based metrics. Let us denote T the template image and I the input image to be tested. If the SSD metric is used, the results matrix of matching M is computed as:

[4] Fileadmin.cs.lth.se

$$M(x,y) = \sum_{x',y'} (T(x',y') - I(x+x',y+y'))^2 \tag{11}$$

If the cross-correlation metric is used, M is computed as:

$$M(x,y) = \sum_{x',y'} T(x',y').I(x+x',y+y') \tag{12}$$

The matching matrix M then allows finding the most probable location of the template image in the input image. Compared to the Haar classifier method, this technique is much simpler but also less efficient, as only one image is used for the template. It is not rotation-invariant, scale-invariant and the template and the input image must have same brightness for good recognition performance. Finally, the Haar classifier is adapted to video analysis and computationally very efficient.

V.3.c. Bag-of-visual-word approach

From the category of feature-based approaches for object recognition, the use of key-points (or points of interest) revealed satisfactory results. Key-points and their local descriptors are very useful for detecting important regions in images that can be studied. Indeed, the repeatability is one of the major characteristics of key-point detection methods. For instance, it is easier to identify and track a key-point on a corner of a table than a key-point in the middle of the table within a uniform colour. These key-points may also have particular parameters, such as scale or orientation invariance. Each key-point is then described (represented) by a vector according to its spatial neighbouring. We first present methods for detecting key-points, then we show the different method for describing the key-points previously detected and we finally present the bag-of-word-approach.

Key-points detection

Given an image I in 2 dimensions, where $I(x,y)$ is the intensity value at point (x,y), four key-points detectors are presented: the Harris, SIFT and SURF key-points, as well as key-points based on mutual information.

Harris key-points

This method has been described by Harris et al. (1988) and focuses on the detection of corners into the images. A corner is a particular point where intensity varies in both space directions, characterized by large variations in x and y. Based on this idea, the second moment matrix A is defined as:

$$A = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix} \tag{13}$$

where I_{xx} and I_{yy} are the second-order partial derivatives with respect to the x and y axis respectively, I_{xy} the first-order partial derivative with respect to the x axis followed by a first-order partial derivative with respect to the y axis. Variations can be determined based on the two eigenvalues of the

matrix A. In the case of only one strong eigenvalue, it is an edge of the image and not a corner. For direct determination of eigenvalues proportions, this formula is used:

$$M = \det(A) - k \times trace^2(A) \tag{14}$$

where $\det(A)$ is the A determinant matrix, and $trace(A)$ the trace of the matrix A. The value of the k parameter is empirically determined, but often set in the literature between 0.04 and 0.15. (x, y) is considered as a key-point when M is superior to a threshold defined according to use conditions.

SIFT key-points

This method has been described by Lowe (1999) and Lowe (2004). It allows the detection of key-points that are invariant to scale, translation, rotation and illumination which make the detection robust. The detection is based on four steps.

The first step allows the creation of a scale space (invariant to scale variations). The input image is convoluted with a Gaussian filter for blurring. This operation allows small details of the image to be blurred or removed.

$$L(x, y, \sigma) = G(x, y, \sigma) I(x, y) \tag{15}$$

where

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{16}$$

σ being the standard deviation of the Gaussian. An increase of the blurring effect appears when the standard deviation of the Gaussian increases. After the creation of multiple burred images at the initial resolution, a downsampling by a factor of 4 is done and same convolutions are performed. It is repeated so that the scale space is finally obtained. The convolved images are grouped by octave (an octave corresponds to doubling the value of σ). **Figure 32** gives an example of scale space using an image extracted from a video of cataract surgery:

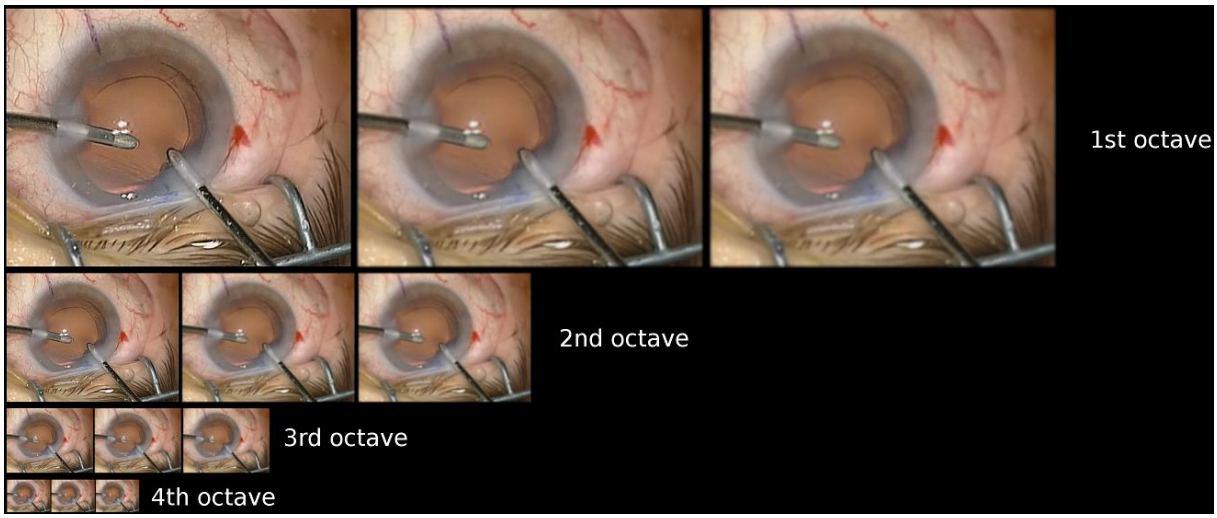


Figure 32 - Scale-space using cataract surgery images.

The second step consists of the creation of Differences of Gaussian (DoG) from the set of images previously created (**Figure 33**). This operation is close to the Laplacian of Gaussian, but instead of computing second-order partial derivatives of images (Laplacian) that are extremely sensitive to noise and that are computationally heavy, the scale space is used. Differences between two successive images within an octave are computed.

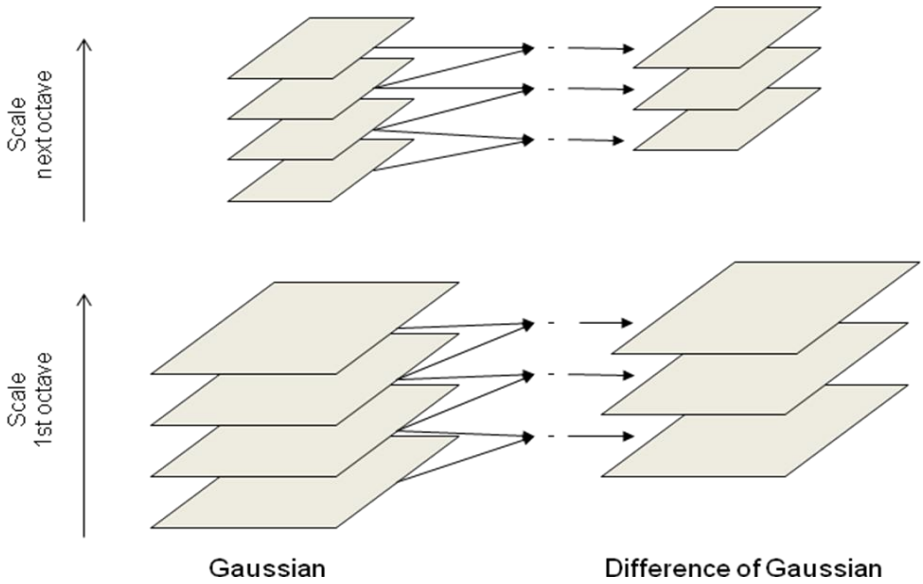


Figure 33 - DoG construction using two scale spaces

In a third step, the maxima and minima positions of the DoG are studied. For each DoG, and for each pixel, the extrema is searched over its neighbors (8 neighbors in 8-connexity) as well as over the previous and next DoG (9 neighbors in 8-connexity), having a total of 26 neighbors to be compared. The key-point is conserved only if it is an extremum over its 26 neighbors.

The fourth step allows the refinement of results by elimination of points being either on edges or on region with light contrast. The method for detecting the Harris key-points is used here but for removing possible points, i.e. when the two gradients (i.e. in both directions) of a point are strong the point is removed. For light contrast points, points are kept only if their values into the DoG are superior to a threshold.

SURF key-points

This method, quite recent and described in Bay et al. (2006), is an improvement of the SIFT key-points in term of robustness and rapidity. In order to accelerate image processing, the fast-hessian approach has been introduced instead of the classical DoG used in the SIFT key-points detector. The fast-hessian is based on the Hessian matrix $H(x,y,\sigma)$, defined as:

$$H(x,y,\sigma)=\begin{pmatrix} L_{xx}(x,y,\sigma) & L_{xy}(x,y,\sigma) \\ L_{yx}(x,y,\sigma) & L_{yy}(x,y,\sigma) \end{pmatrix} \tag{17}$$

where $L_{xx}(x,y,\sigma)$ and $L_{yy}(x,y,\sigma)$ are the convolutions of the second-order partial derivative of $L(x,y,\sigma)$, with respect to the x and y axis respectively. Similarly, $L_{xy}(x,y,\sigma)$ and $L_{yx}(x,y,\sigma)$ are defined as the first-order partial derivative with respect to the x axis followed by a first-order partial derivative with respect to the y axis. The determinant of $H(x,y,\sigma)$ is therefore:

$$\det(H(x,y,\sigma))=\sigma^2(L_{xx}(x,y,\sigma)L_{yy}(x,y,\sigma)-L^2{}_{xy}(x,y,\sigma)) \tag{18}$$

By searching for local maxima of this determinant, a list of K points that are associated with a value of σ is first established. The number of points depends on the scale space (defined for the SIFT key-points) and on a threshold for local maxima. A way to avoid heavy computation time is to approximate the second-order partial derivative of Gaussian filter (with σ=1.2) with box filter of 3x3 (**Figure 34**), denoted D_{xx} D_{yy} and D_{xy}.

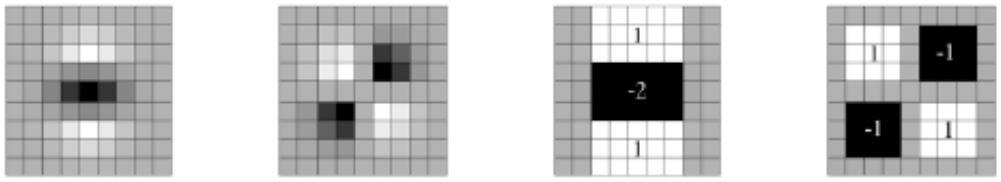


Figure 34 - Gaussian approximation for the SURF method[5]. Left to right: the (discretised and cropped) Gaussian second order partial derivatives in y-direction and xy-direction, and the approximations using box-filters.

In the case of approximations using box-filters, a new determinant of the Hessian matrix $H(x,y,\sigma)$ is computed:

$$\det(H_{approx}(x,y,\sigma))=D_{xx}(x,y,\sigma)D_{yy}(x,y,\sigma)-(0.9D_{xy}(x,y,\sigma))^2 \tag{19}$$

[5] Bay et al. (2006)

with the parameter 0.9 obtained using:

$$\frac{\left|L_{xy}(1.2)\right|_F}{\left|L_{xx}(1.2)\right|_F}\frac{\left|D_{xx}(9)\right|_F}{\left|D_{xy}(9)\right|_F} = 0.912.. \approx 0.9 \tag{20}$$

and $\left|x\right|_F$ being the Frobenius norm (Golub and Van Loan, 1996).

Mutual information key-points

This method, described by Dame and Marchand (2009), proposes the detection of key-points using the entropy of the image. A metric based on mutual information is used instead of a sum-square-difference (SSD) metric because mutual information is less sensitive to illumination variations and some non-linear transformations. Mutual information is defined as the quantity of information shared between two random variables (here two images). The method is also based on the Hessian matrix, which is based in this case on the joint probability:

$$\begin{aligned} H &= \frac{\partial^2 MI(T_0, w(T_t, p))}{\partial p^2} \\ &= \sum_{r,t} \frac{\partial p_{r,t}}{\partial p}^T \frac{\partial p_{r,t}}{\partial p}\left(\frac{1}{p_{r,t}} - \frac{1}{p_r}\right) \end{aligned} \tag{21}$$

A circular Gaussian centred on key-points is used as a weighted function in order to compute the joint probabilities:

$$p_{rt}(r,t,p) = \frac{1}{N_x}\sum_x \pi(x)h(r,t,p) \tag{22}$$

with

$$\pi(x) = \frac{1}{\sigma\sqrt{2\pi}}e\left(-\frac{(x - x_c)^T(x - x_c)}{2\sigma^2}\right) \tag{23}$$

x_c is the coordinate vectors of the center of the window and σ^2 is the variance of the selected Gaussian.
In presence of key-points, this matrix contains strong eigenvalues. These eigenvalues are computed using a Singular Value Decomposition (SVD).

Key-points descriptors

Six key-point descriptors are presented here for 2D images: The Harris, SIFT, SURF, GLOH, SIFT-Rank and PCA-SIFT descriptors. Others extensions or improvements of these methods have been recently proposed, such as the Histogram of Oriented Gradient (HOG) or the Local Energy based Shape Histogram (LESH) descriptors, but the 6 following methods allow having a good overview of current local key-point descriptor methods.

Harris descriptors

A region of size 16x16 is chosen around the key-point, and then this region is subdivided into sub-regions of size 4x4. For each sub-region, a vector (histogram) composed of eight parameters is computed, each of the parameter representing an orientation according to the Freeman representation (**Figure 35**). The final histogram will be of size (16x16)/(4x4)x8=16x8=128. In order to fill each sub-region, the magnitude and the orientation of the gradient are computed for each point. Histograms will be filled from these values and using the Freeman representation. In each point, a projection of the gradient vector $V = (dx, dy)$ on two nearest directions is performed. The histogram is then incremented with both projection values. Magnitude values are normalized in order to obtain the maximum value after the projection equal to 1.

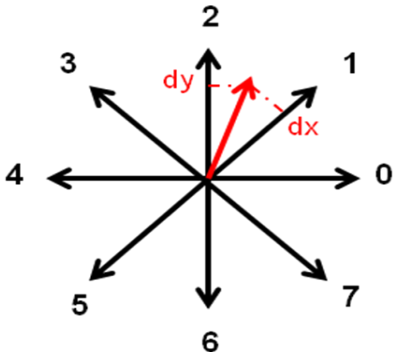


Figure 35 - Freeman representation.

SIFT descriptors

Similar to the Harris descriptors, a region of size 16x16 is considered around the key-point, and then this region is subdivided into sub-regions of size 4x4. For each sub-region, an histogram of 8 parameters is computed according to the following representation:

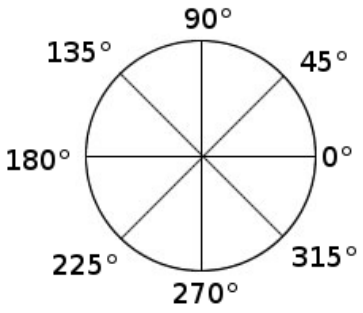


Figure 36 - Representation of orientations for SIFT descriptors.

To fill the histogram of sub-regions, the gradient magnitudes and orientations are also computed. The histogram is incremented according to the most probable orientation. This value is of course dependant of the magnitude and of the distance of the key-point. A weight is applied using a Gaussian

function where σ is equal to ½ of the size of the region. Once the 128 values are obtained, normalization is performed in order to reduce illumination change effects. Here is a simplified representation:

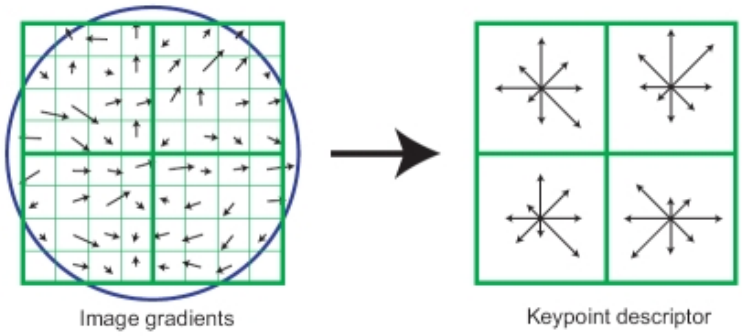


Figure 37 - Simplified representation of the SIFT descriptor method[6]

SURF descriptors

The same regions and sub-regions are computed, but with 4 parameters per sub-regions instead of 8 used for Harris and SIFT descriptors. Haar wavelets are used to compute values for the vector. Let us denote dx the response of the Haar wavelet on horizontal direction, and dy the response in vertical direction. Both directions are defined according to the key-point orientation. For each sub-region, the vector v is computed as:

$$v = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_x|) \tag{24}$$

SIFT-Rank descriptors

This ranking method is very advantageous when descriptors are of big size (e.g. SIFT). The idea is to re-organize descriptor vector values in an ordinal manner. The initial descriptor vector is transformed into a rank vector. In order to be able to compare two vectors, methods like Squared Euclidean Distance, Spearman correlation coefficient or Kendall coefficient can be used.

PCA-SIFT descriptors

This method has been described by Ke and Sukthankar (2004) and is based on SIFT descriptors. A region of size 41x41 around the key-point is extracted. A rotation is then performed so that its principal direction is aligned with a direction from the canonical base (e.g. vertical). From this initial step, PCA-SIFT can be resumed into 3 steps. The first step, which has to be done off-line, is the pre-computation of an eigen-vectors space to express gradients of the local region. Around 21000 key-points have to be computed with the SIFT method, and vectors of 3042 elements are created from the local horizontal and vertical gradients of each region. This vector is normalized in order to minimize illumination variations. A PCA is then performed on the covariance matrix of these values. The matrix having the n best eigen-vectors is kept and will correspond to the eigen-vectors space. A projection of

[6] cs.washington.edu

the gradient vectors in the eigen-vectors space is finally performed to obtain the descriptor vector. The size of the final descriptor vector is therefore inferior to the one obtain using the SIFT method. An Euclidean distance can be used between two PCA-SIFT descriptors in order to determine if both vectors belong to the same key-point in different images.

GLOH descriptors

Gradient Location-Orientation Histogram (GLOH) descriptors have been described by Mikolajczyk and Schmid (2005) and are extension of SIFT descriptors. In addition to the 8 angular parameters of the SIFT descriptors, 3 radials directions are added considering that no subdivision is performed in the smallest radius zone (**Figure 38**). Histograms of sub-regions have therefore a size of 17 instead of 8, which is too important. A PCA is finally done to reduce histograms size to 128 or even 64.

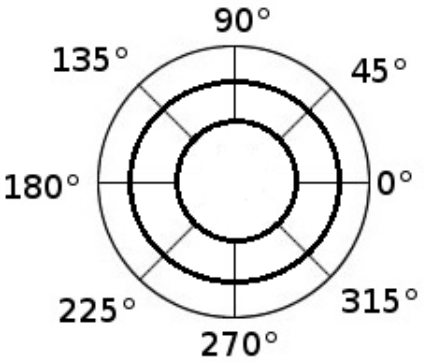


Figure 38 - Representation of orientations for GLOH descriptors.

Bag-of-visual-word algorithm

For whole-image categorization tasks, or for recognition of objects in images based on local information, bag-of-visual-words (BVW) representations, which represent an image as an orderless collection of local features, have recently demonstrated impressive levels of performance along with relative simplicity of use. The idea of BVW is to treat images as loose collections of independent patches, sampling a representative set of patches from the image, evaluating a visual descriptor vector for each patch independently, and using the resulting distribution of samples in descriptor space as a characterization of the image. Given the occurrence histograms of positive and negative regions of a training database, a classifier can then be trained. Three steps can be defined, and a fourth one corresponding to the creation of a vocabulary can be additionally considered.

- ✓ Step n°1: The methods previously described are used here to extract key-points. These key-points are also called word. For each image, a set of key-point is obtained. **Figure 39** shows an example for one cataract surgery image.

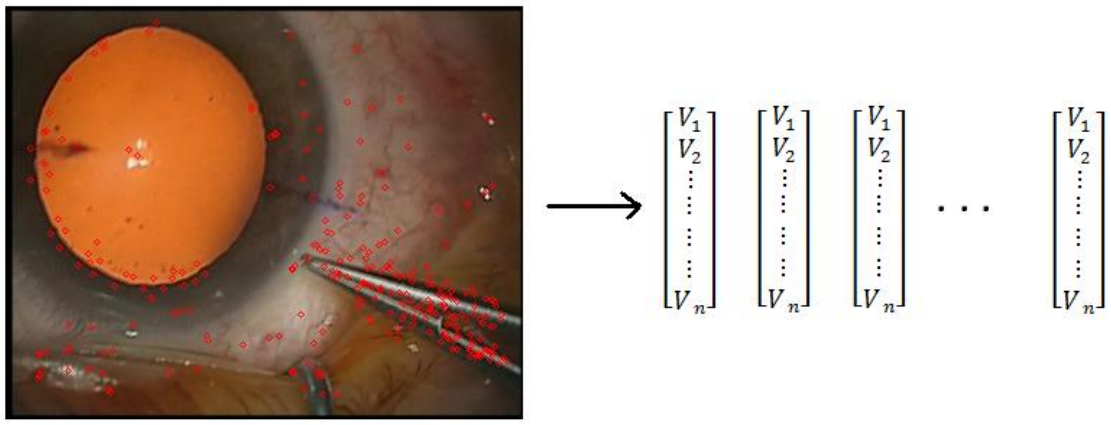


Figure 39 - Representation of key-points obtained using the SIFT method over one entire cataract image.

- ✓ Step n°2: After detection, a key-point is then described as a local, rectangular or circular, patch of the image and is represented in a formal way. Each key-point is thus represented by a descriptor vector whose length is variable and highly correlated to the chosen descriptor method. The different methods previously enumerated (**Figure 39**) can be applied here. After this step, an image is represented by a collect of vectors of same dimension (e.g. 128 for SIFT).

- ✓ Step n°3: A training database has to be created in which the object has to be incorporate. For each image of the database, a set of descriptor vectors is created. Using all vectors, a global clustering (using a k-means algorithm) is performed, and the number of cluster (i.e. word) to be kept has to be set by the user according to the context. Lastly, each image of the database is transformed into a histogram where for each word is associated its number of occurrence found in the image (a bag of keypoints is expressed as a histogram recounting the number of occurrences of each particular pattern in an image, **Figure 40**). During this phase, each descriptor vector is associated to the closest word in the vocabulary in term of a metric (often an Euclidean distance). **Figure 40** shows a visual representation using a cataract image again.

- ✓ Step n°4: it's the final step of the method, which involves the use of a vocabulary of the object to be identified. From the descriptor vectors of the input image and a supervised classifier (e.g. SVM) trained over the image database, a class is assigned to the input image. For this step, a class has to be first assigned to each image of the database according to the goal of the study.

This technique, used for instance in the context of medical imaging by Andre et al. (2009), have been widely used for object recognition in image or videos and will be used in this thesis for recognizing local information such as local texture or objects.

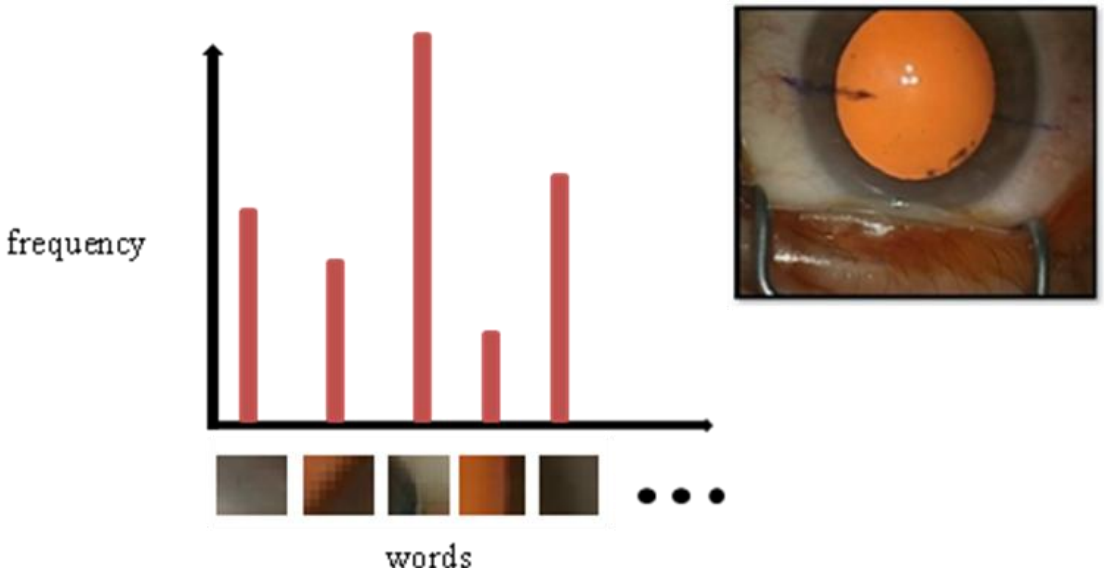


Figure 40 - Representation of an image using a histogram of words from a vocabulary.

V.4. Temporal features

As previously outlined, it could be of interest to extend the concept of spatial features to the temporal domain. In this subsection, we will investigate the use of temporal features as possible information that can be extracted from videos. Two main possibilities exist for such approach. The first one is the possibility of extracting spatio-temporal features that can enrich image signatures by integrating complementary features. The second possibility is the detection of movement in videos, including object tracking.

V.4.a. Spatio-temporal features

Spatio-temporal features allow characterizing objects that are moving in the video. They belong to the family of low-level features but are still very robust. The method described by Laptev and Lindeberg (2006) is presented here. This technique is an extension of the Harris method with a temporal component. The idea is to start from a classical Hessian matrix where the time parameter has been added. For an image $I(x, y)$:

$$H(x, y, t) = \begin{pmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial x \partial t} \\ \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial y^2} & \frac{\partial^2 I}{\partial y \partial t} \\ \frac{\partial^2 I}{\partial x \partial t} & \frac{\partial^2 I}{\partial y \partial t} & \frac{\partial^2 I}{\partial t^2} \end{pmatrix} \quad (25)$$

For the second derivatives, convolutions with Gaussian are done. A spatial Gaussian as long as a temporal Gaussian are used with their dedicated parameters:

$$\frac{\partial g_{\sigma_s}(x,y)}{\partial x} \otimes \left(\frac{\partial g_{\sigma_t}(t)}{\partial t} \otimes I(x,y,t) \right) \tag{26}$$

By following the same principle than for the Harris method, spatio-temporal key-points are points where eigenvalues are high in every direction, i.e. when $R(x,y,t)$ is superior to a threshold:

$$R(x,y,t) = \det(H(x,y,t) - k \times trace(H(x,y,t))^3 \tag{27}$$

V.4.b. Object tracking

Object tracking techniques have various applications, from human/machine interfaces to surveillance, video compression, augmented reality or medical imaging. It's a highly time computational challenge and the complexity often increases when object recognition techniques are associated. It can be very difficult when the object has a rapid movement compared to the video flow. One way to detect movements in video is the use of optical flow (Beauchemin and Barron, 1995; Fleet and Weiss., 2006). This technique allows the detection of visible and significant movements of structures (objects, surfaces, corners), induces either by a displacement between the camera and the surgical scene or by the own object displacement. The algorithm computes a displacement for each pixel between a video frame at time t and a frame at time $t+\delta t$. In the case of a $2D+t$ video, with a pixel (x,y) of intensity $I(x,y)$ at time t: the pixel will move of δx, δy, δt, which can be written by:

$$I(x_t, y_t) = I(x_{t-\delta t} + \delta x, y_{t-\delta t} + \delta y) \tag{28}$$

Multiples algorithms have been proposed to compute the optical flow:

- ✓ Block-matching algorithm: the goal is to find a matching between a block of the frame at time t and the frame at time $t+1$. The positions of both blocks then allow the computation of a displacement vector.

- ✓ Lucas-Kanade algorithm (Lucas and Kanade, 1981): the algorithm is based on three assumptions: the intensity conservation between two consecutives frames, the temporal persistence and the spatial coherency (neighbouring pixels follow a similar displacement). Windows of 5x5 around each pixel are defined and a least-square minimization is computed to solve the problem and find the displacement.

- ✓ Horn-Schunck method (Horn and Schunk, 1981): the estimation of the optical flow is performed by incorporating a constraint of smoothness. The main assumption of this method is the spatial coherency. The optical flow is formulated as a global energy function, and its minimization is performed using the Euler-Lagrange equations (Fox, 1987).

One limitation of such local methods appears in the case of big displacements. In such cases, the new position can be out of the local window defined for the estimation. One solution to this problem is the use of multi-resolution that permits to detect a big displacement at the top of the pyramid (low resolution) and then refines the results along the pyramid. Without multi-resolution, this technique remains computationally heavy, and the complexity increases when object recognition techniques are added. Moreover, it turns out to be difficult when objects have fast movements.

V.5. Methods

In this section, we present how we extended the previous framework for incorporating local features, temporal features and object detection and recognition in order to improve the performances of surgical phases recognition from video images.

V.5.a. Framework presentation

The proposed framework (**Figure 41**) was created to be adapted and adjusted, if needed, to different types of surgical procedures. The idea of this extended framework is first to manually defined visual cues that can be helpful for discriminating high-level tasks. These visual cues are the important information of the surgery that allows an observer to segment the surgical workflow into phases. It can be, for instance, the presence of a surgical tool, a particular colour that appears in the surgical scene or even a particular texture. These visual cues were automatically detected using the local features presented in the previous subsections or using the low-level image features presented in the previous chapter. The recognition of these visual cues, as statically extracted at each time step, allows the creation of a feature vector for each video frame that can be seen as a particular frame signature composed of some high-level information. The sequences of frame signatures are then analysed using time-series analysis to recognize surgical phases. The DTW algorithm and the HMM algorithm were used to recognize surgical phases.

Compared to traditional video understanding algorithms, this framework extracts generic application-dependant visual cues. The combination of image-based analysis and time series classification enables high recognition rates to be achieved. We validated each part of the framework with both data-sets through various cross-validation studies, and finally compared global recognition rates obtained using of the DTW approach and the HMM classification.

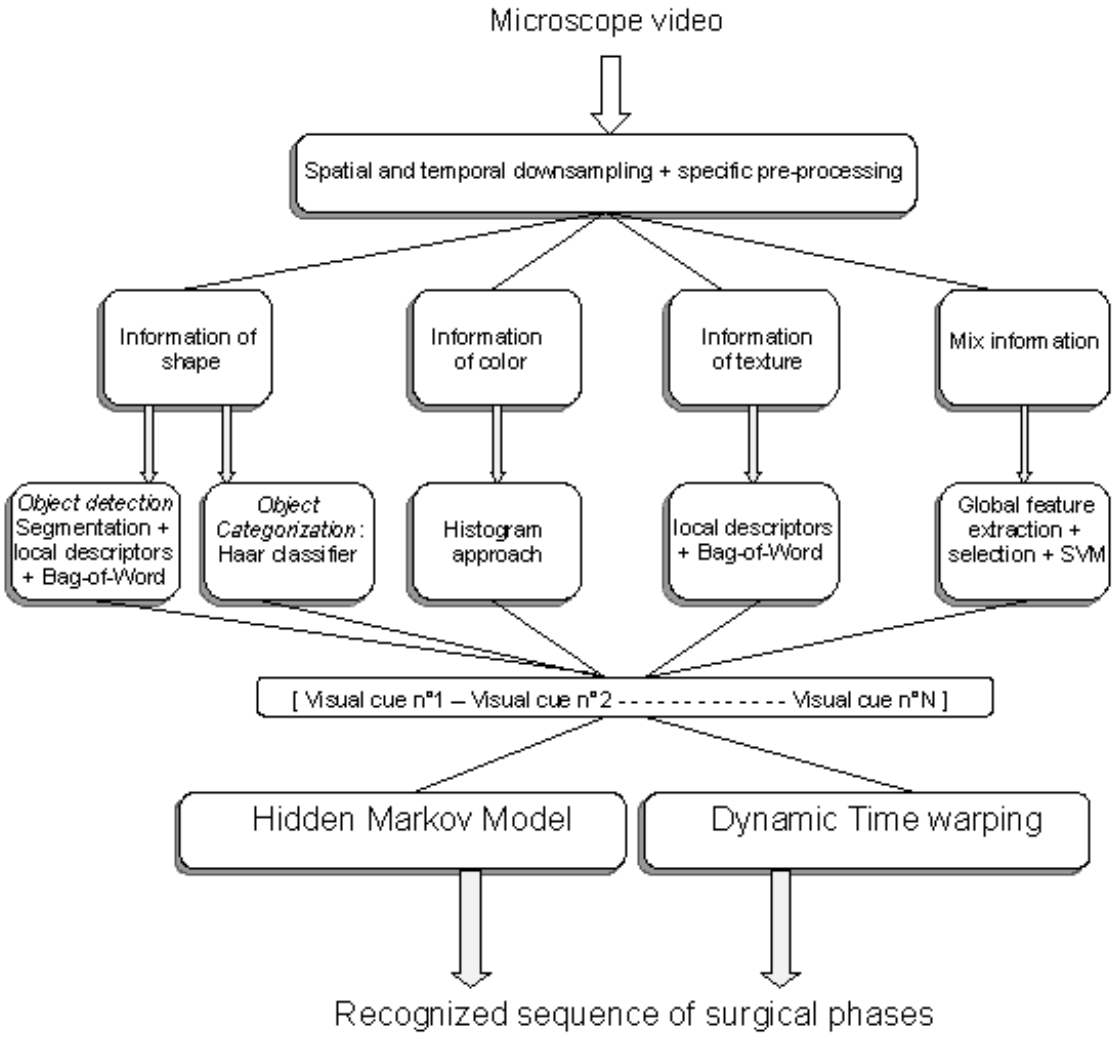


Figure 41 - Framework of the recognition system.

V.5.b. Pre-processing steps

Pre-processing steps were applied to both datasets separately, but with more efforts on the cataract one. Indeed, the microscope in cataract surgery has the same focus as well as the same magnification values all along the surgery. The only microscope parameter that varies in time is its displacement, which makes the pupil not always perfectly centred. This parameter allows us to automatically apply image processing operations without modifying any parameters in the pre-processing algorithms. For the pituitary dataset, the zoom is always changing, removing the possibility of using any segmentation processes. That's the reason why, after a down-sampling performed on both datasets, segmentation and connected component detection were applied to cataract surgery videos only, while no further investigations were made for the pituitary video pre-processing.

Down-sampling

For this second framework, the same time and spatial down-sampling than the previous study was performed for the pituitary surgery videos, i.e. 1 fps (1Hz) and a spatial down-sampling by a factor of 4. For the cataract surgery videos, and similarly to the pre-processing of the first dataset, we down-sampled the videos to 1 fps and performed a spatial down-sampling by a factor of 8 using a Gaussian kernel.

Pupil Segmentation

For the dataset of cataract surgery, some visual cues are identifiable inside the pupil only. The regions around the pupil may therefore bias the detection and a preliminary segmentation step is needed to ensure good detection. Detecting this ROI will allow the retrieval of more specific visual cues, consequently improving phase recognition. The pupil segmentation procedure can be divided into three steps and is based on the colour difference between the pupil and the remaining eye regions.

- ✓ The first step allows the creation of an outline mask from the input image transformed into the YUV colour space. Within this first step, smoothing (Gaussian filter 5x5), thresholding (set to 127 over 256 greyscale values) and morphological operations (as presented in subsection V.2.b.) were performed to the input image, obtaining a binary mask (**Figure 42**, middle left).

- ✓ Using this binary mask, the second step consists in determining circles through the image using the Hough (Hough, 1959) transform (**Figure 42**, middle right). A choice between the different circles has to be done, based on a reference diameter.
- ✓ **Figure 43** shows an example of multiple circles found using the Hough transform.

- ✓ The third step can be considered as a normalisation step. As no zooms are performed by surgeons during the intervention, all pupil have sensibly the same diameters. The circle that was kept is therefore re-adjusted using a reference radius and applied to the input image. The reference radius was defined according to the average of 100 radiuses computed from manually segmented images. Following this procedure, the ROI around the patient pupil can be retrieved (**Figure 42**, right).

Figure 42 - Different steps of the pupil segmentation. From left to right: input image, 1st step: creation of the mask, 2nd step: Hough transform computation, 3rd step: final segmentation of the pupil.

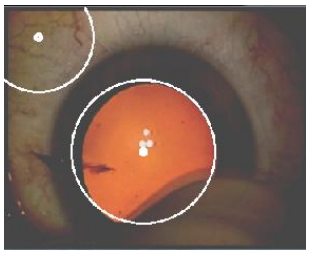


Figure 43 - Illustration of multiple Hough circles found in an image.

Sometimes, incomplete circle outlines in the mask may occur, leading to Hough circle detection failure. To tackle this problem, we proposed a second method that has been also implemented. An iterative search was performed on the binary mask to identify the most probable circular zone. This search was based both on pixel counting and circle radius assumption. Even if this alternative method showed satisfactory results, we didn't integrate it into the framework. Indeed, it could have been used as an alternative method when the first method was not accurate, but no reliable solutions were found to detect when errors occurred in the Hough circle detection, which didn't permit to use this second method.

Connected components detection

The extended recognition framework aims at integrating information about tools and zones. For analyzing this type of information, the first step consists in extracting ROIs using connected components for being able to detect surgical tools. The goal was to create as many ROIs as surgical tools in the image. The better the ROIs around the tools are, the better the identification will be. As this step is also mandatory for analyzing the zone where tools are used, we consider it as a pre-processing step.

First, the input image is transformed into a binary mask, as performed for the pupil detection step. Image processing operations were applied: Gaussian smoothing, Laplacian, threshold and a dilatation. The dilatation increases the size of the detected zones, which compensates the size decrease induced by the primary processing (**Figure 44**).

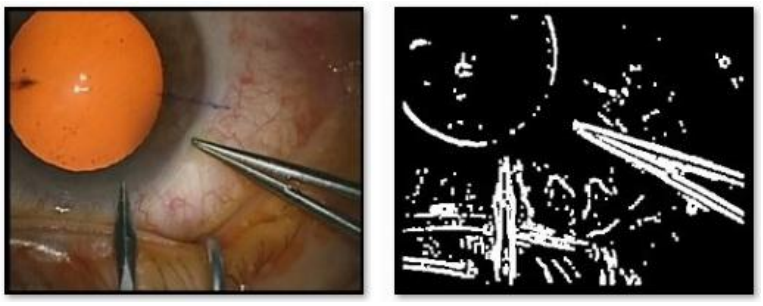


Figure 44 - Illustration of the binary mask for the creation of the ROIs.

Then, the mask is refined by applying connected component operation, as presented in subsection V.2.c., in order to remove artifacts. By applying a connected component method to the mask, we were able to detect and remove all small connected components which were assumed to be noise. We used an 8-connexity metric for the processing, and the threshold was empirically defined (**Figure 45**).

Figure 45 - Illustration of the connected components operation.

The mask now only contains important ROIs resulting from strong edges of the input image. In the specific case of cataract procedures, more than 2 surgical tools can't be present at the same time in an image (one per hand). That's the reason why we retrieved the two largest (in term of number of pixels) remaining connected components respectively and created a mask for each one. At this stage, these selected connected components are very likely to be the instruments. By applying these masks to the input image, we obtained two different images, each one with only a ROI of the input image (**Figure 46**, **Figure 47**).

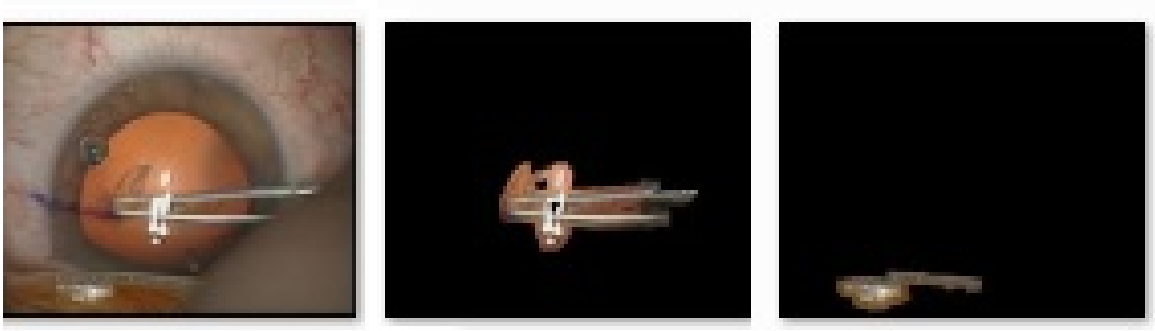


Figure 46 - 1st illustration of the creation process of the two ROIs. From left to right: input image, ROI n°1 corresponding to the first connected component, ROI n°2 corresponding to the second connected component.

Figure 47 - 2nd illustration of the creation process of the two ROIs. From left to right: input image, ROI n°1 corresponding to the first connected component, ROI n°2 corresponding to the second connected component.

V.5.c. Application-dependant visual cues

Five sub-systems based on different image processing tools were implemented. Each of these sub-systems is related to one type of visual cue. Visual cues recognizable through colour were detected with simple histogram intersection. For shape-oriented visual cues such as object recognition, a Haar classifier was trained. For texture-oriented visual cues, we used a bag-of-words approach using local descriptors, and finally for all other visual cues we used a conventional image classification approach including a feature extraction process, a feature selection process and a supervised classification with SVM. In all cases, the features were considered to be representative of the appearance of the cues to be recognized.

Colour-oriented visual cues

Colour histograms have a long history as a method for image description, and can also be used for identifying colour shade through images. We used here the principle of histogram intersection to extract colour-oriented visual cues, by creating a training image database composed of positive and negative images. Two complementary colour spaces (Smeulders et al., 2000) were extracted: RGB space (3 x 16 bins) along with Hue (32 bins) and Saturation (32 bins) from HSV space. For classifying visual cues, we used a KNN classifier with the correlation distance to compare histograms composed of feature vectors.

Texture-oriented visual cues

Based on the key-points detection and description, a BVW approach (as presented in subsection V.3.c) is used here to identify and classify particular texture in the image. The detection of key-points is performed only on the pupil after preliminary segmentation. The final supervised classification step was performed using a KNN algorithm. Global image texture can be categorized using this approach. **Figure 48** shows 2 examples of the extraction of local features after the pupil segmentation step.

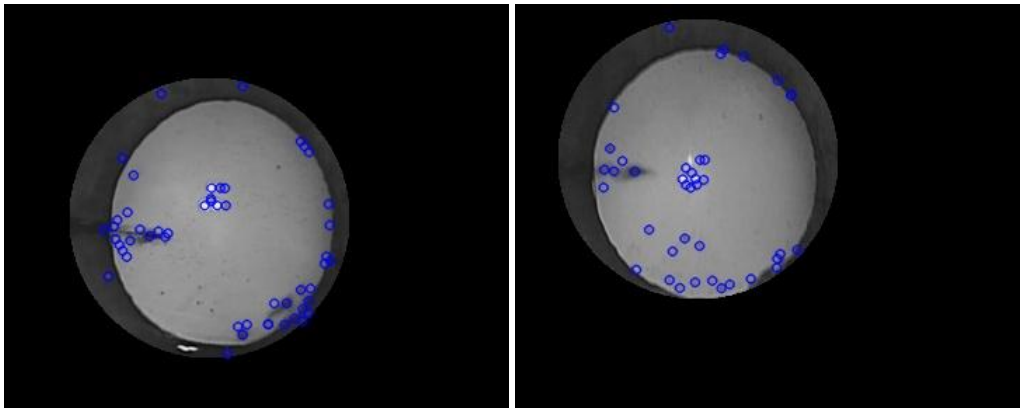


Figure 48 - SIFT features detected on 2 images and shown as blue circle.

Shape-oriented visual cues - surgical tools categorization or detection

The detection of tools in the surgical layout is a vital piece of information to access finer details in surgical process analysis. The main limitation is that instruments frequently have similar shapes and are therefore very difficult to recognize through image-based analysis only. Two methods were thus implemented: one for categorizing highly recognizable surgical tools, and one for detecting the presence or not of any tools.

We implemented a Haar classifier, as presented in subsection V.3.a, for recognizing and categorizing recognizable surgical tools with strong edges. We chose this approach for computational reasons, getting a robust method that minimizes computation time while achieving high detection accuracy. This algorithm is known to work well for rigid objects, so we applied it in the case of representative surgical tools categorization.

A second method was implemented for tools that have similar shapes and low edges. This required the use of local information. We used a BVW approach using a preliminary ROIs segmentation step (explained in subsection V.5.b) compared to texture-oriented visual cues where the detection of key-points was performed only on the pupil. Then, for the description step, the aim was to provide a robust and reproducible method for describing the ROIs that have been isolated by the segmentation step. The final supervised classification step was also performed using a KNN algorithm. **Figure 49** shows an example of the extraction of local features for the input image of **Figure 46**. If the same instrument appears with various scales and orientations, we will be able to extract the same feature points with the same descriptors. Using this approach, two classes can be defined: one class including all surgical tools, and one class including the background.

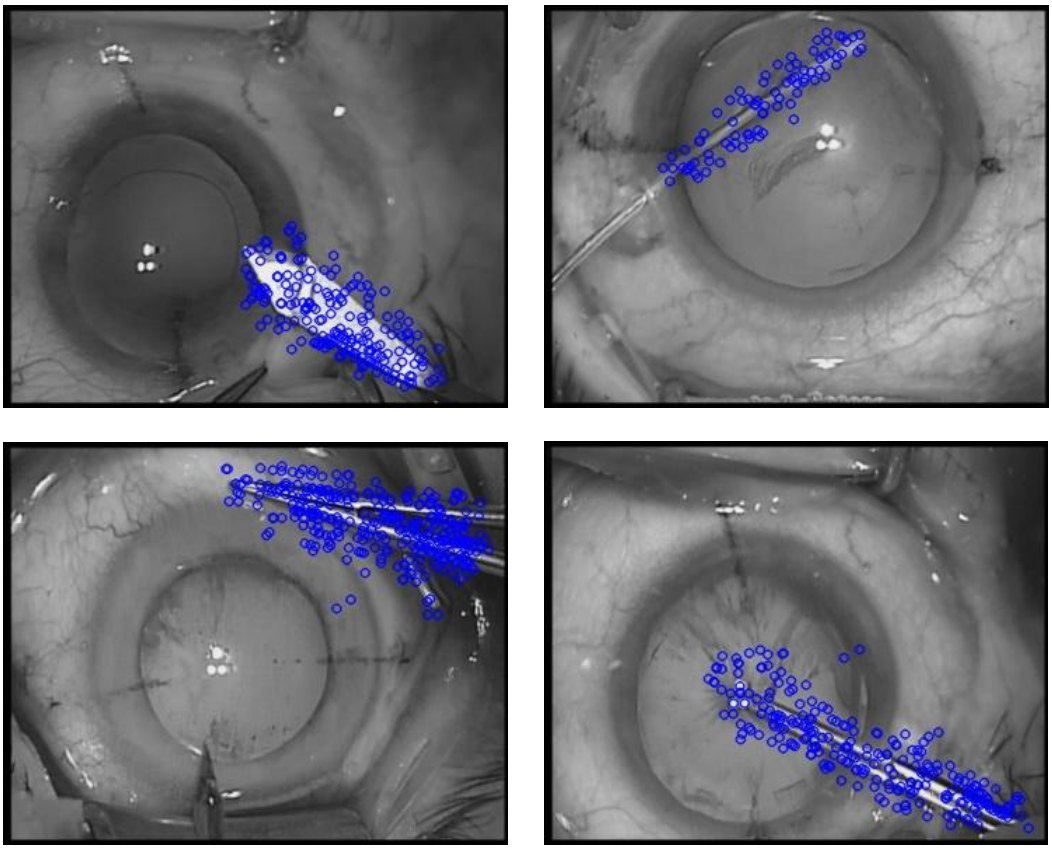


Figure 49 - SURF features detected on different ROIs and shown as blue circles.

Alternative method

This approach is the one presented in Chapter IV. It was created to be used in particular cases where the visual cues are not detectable through only texture, colour or shape analysis. As explained in the previous chapter, it combines complementary features that allow creating complex image signatures that can then be classified using state-of-the-art supervised classification techniques. **Table 9** gives all parameter values used in all of the 6 visual cues detection methods.

Table 9 - Parameters of the classification algorithms used for extracting visual cues.

Type of visual cues	Algorithm	Parameters
Colour-oriented	Color histogram intersection	*Type of color space*: RGB, HSV *Classifier*: KNN *Distance*: correlation
Texture-oriented	BVW approach	*Classifier*: SVM with Gaussian kernel *Key-points detector*: SIFT/SURF/Harris/STAR *Key-points descriptor*: SURF *Codebook generation*: KNN
Instrument categorization	Viola-Jones approach	*Features*: Haar-like rectangular *Number negative images*: 2000 *Number positive images*: 500
Detection of other instruments	BVW approach	*Classifier*: SVM with Gaussian kernel *Key-points detector*: SIFT/SURF/Harris/STAR *Key-points descriptors*: SURF *Codebook generation*: KNN
Alternative method	Global features classification	*Spatial features*: RGB, HSV spaces, Haralick descriptors, DCT, spatial moments *Wrapper method*: RFE-SVM *Filter method*: MI *Classifier*: SVM with Gaussian kernel

V.5.d. Visual cues definition and extraction

The purpose of this step was first to define relevant binary cues from the microscope images that can differentiate surgical phases. In other words, the surgeon was asked to define visual information that was linked to each specific phase. It was requested that such binary cues should be easily identifiable through image-based analysis, and that each of them should take two values only (i.e. binary signal) be taken. This could be, for instance, the presence/absence of a specific surgical object in the operating field, or the microscope view (zoom or not). When the visual cues are defined, each phase must have different reference signature compared to its precedent and consecutive phase so that the time-series analysis could explicitly differentiate them. After the definition of the visual cues, one of the five methods presented in subsection V.5.c has to be chosen for each visual cue.

Pituitary surgery videos

Four discriminant pieces of binary cues were defined for this surgery: global-zoom views, presence/absence of nose retractors, presence/absence of the column of nose and presence/absence of the compress. Experiments have been conducted in order to choose the best type of algorithm for detecting these visual cues. The more coherent choice would be to use a Haar classifier for the detection of nose retractors due to its particular shape, a colour histogram for the presence of the compress due to its white colour and the alternative method for the presence of the column of noise

and for global-views zoom that are visually not identifiable through only shape, texture or colour. This choice was validated and details of these experiments can be seen on subsection V.5.f. Relations between these visual cues and the surgical phases are shown on **Table 10**. In this surgery, phase n°1 and phase n°5 have for instance same image signatures, but they were all different between consecutive phases.

Phase	1-Nasal Incision	2-Nose retractor installation	3-Access to the tumour + tumour removal	4-Column of nose replacement	5-Suturing	6-Nose compress installation
Global-zoom View	False	False	True	True	False	False
Presence-absence nose retractors	False	True	True	True	False	False
Presence-absence column of nose	False	False	False	True	False	False
Presence-absence compress	False	False	False	False	False	True

Table 10 - Relations between the surgical phases and the binary cues for the pituitary data-set.

Cataract surgery videos

Six discriminant pieces of binary cues were defined here. The pupil colour range, defined as being orange or black, was extracted using a preliminary segmentation of the pupil as explained in subsection V.5.b, along with a colour histogram analysis. Also analysing only the pupil after the segmentation step, the global aspect of the lens (defined as parcelled out or not) was recognized using the BVW approach only on the pupil with local spatial descriptors, as being a texture-oriented visual cue. The presence of antiseptic, recognizable by virtue of its specific colour, was detected using colour histogram analysis, but on the entire image without any pre-processing step of segmentation. Concerning the detection of surgical instruments, only one had a characteristic shape, the knife. We trained a Haar classifier for detecting this specific surgical tool. All other instruments have very similar shapes and are very difficult to categorize. For this reason, we chose to use the BVW approach using the preliminary step of ROIs segmentation. Lastly, the IOL instrument was not readily identifiable through only colour or shape analysis and we chose a classical approach using many spatial features along with a SVM classification to detect this visual cue. The different choices of algorithms have been validated. Please see subsection V.5.f for details. Similar to the pituitary data-set, relations between these visual cues and the surgical phases are shown on **Table 11**. Similar to the other dataset, some image signatures are not unique (e.g. phase n°1 and phase n°8).

Phase	1-Preparation	2-Betadine injection	3-Corneal incision + viscoelastic injection	4-Hydrodissection + capsulorhexis	5-Phacoemulsification	6-Irrigation + aspiration of remanescent lens	7-Implantation of the artificial IOL	8-Adjustment of the IOL + wound sealing
Pupil colour Range	False	False	False	False	True	False	False	False
Presence-absence antiseptic	False	True	True	False	False	False	False	False
Cataract aspect	False	False	False	False	True	True	False	False
Presence-absence knife	False	False	True	True	False	False	False	False
Presence-absence IOL instrument	False	False	False	False	False	False	True	False
Presence-absence surgical tools	True	False	True	True	True	True	True	True

Table 11 - Relations between the surgical phases and the binary cues for the cataract data-set.

V.5.e. Time-series modelling

Once every visual cue of a particular surgery has been defined and detected, creating a semantic signature composed of binary values for each frame, the sequences of frame signatures (i.e. the time series) must be classified using appropriate methods. From the wide choice of time-series analysis methods presented in subsection V.5.e, two different approaches were tested here: the HMM modelling and the classification by DTW alignment. This choice has been motivated by works of Blum et al. (2010) or Padoy et al. (2010) that stated that these two algorithms were very efficient for such time-series analysis. Compared to the DBN that is much more complex, the HMM presents the advantage to be model a single discrete random variable which is well adapted to our application.

HMM

For the first approach, we modelled the sequential aspects of the surgical procedure broke down into phases with a first-order HMM (left-right HMM). According to subsection V.1.a, an HMM is defined by a 5-tuples (S,O,Π,A,B) representing the states, the vocabulary, the initial state probabilities, the state transition probabilities and the output probabilities. States were represented by the surgical phases that generated the HMM (**Figure 50**). Indeed, for the pituitary dataset the HMM was composed of 6 states, and for the cataract dataset the HMM was composed of 8 states. Outputs of the visual cues recognition were treated as observations for the HMM that create the vocabulary. The initial state probability was defined by a single value on the first state, as both surgeries always begin by the same phase. Transition probabilities from one state to its consecutive state were computed for each training video, and then averaged. If we set one probability to α, the probability of remaining in the same state is then $1-\alpha$. Transition probabilities of both models were low because of the low frame sampling rate. Finally, output probabilities were computed as the probability of having an observation in a specific state. Training videos were applied to the supervised classification for extracting binary cues and output probabilities were obtained by manually counting the number of occurrence for each state.

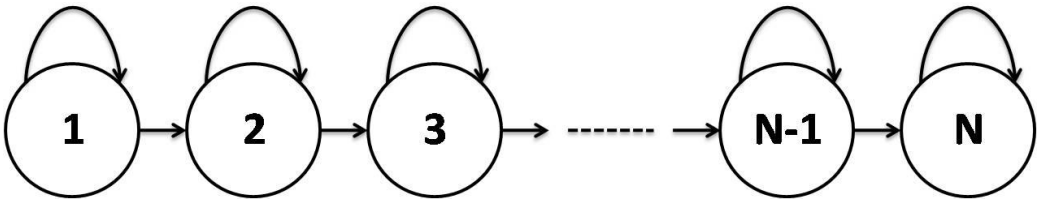


Figure 50 - Left-right HMM used for the analysis.

DTW

In order to use the DTW algorithm as a classifier, an average surgery has to be created based on a learning dataset using the method described by Wang and Gasser (1997). Each "query" surgery is first processed in order to extract visual cues, and then the sequence of image signature is introduced in the DTW algorithm to be compared to the average surgery. Once warped, the phases of the average surgery are transposed to the unknown surgery in a supervised way. For surgery modelling, we used the Itakura parallelogram, presented in subsection V.1.c that adds a constraint to the warping path. This prevents the warping path from straying too far away from the diagonal path. Moreover, considering that we are using binary vectors, we used the Hamming distance for the local distance measure.

V.5.f. Validation studies

Images and videos databases

The majority of algorithms used in the framework were based on machine learning techniques requiring a training stage. We therefore decided to validate each part of the framework through cross-validation studies. This technique is very efficient to have an estimate of the accuracy of a recognition system, but requires a complete labelling of the data-set. Initial indexing was therefore performed by surgeons for each video of both data-sets. In order to be less time-consuming, and knowing that phases always appear on a sequential way, surgeons only defined phase's transitions which considerably reduced the labelling time. The same work has been done for the labelling of all visual cues. Unfortunately, this work was very time consuming and no good strategies were found to reduce time. After deep explanations from surgeons on how the labelling of visual cues should be performed, we therefore did it on our own.

We created two types of databases for each surgery: one image database for the assessment of visual cue detection and one video database along with their corresponding frames from the image database for the assessment of the entire framework. For the image database of pituitary surgeries, 200 frames were randomly extracted from each video, resulting in a database of 3000 labelled images. For the images database of the cataract surgeries, 100 frames were randomly extracted from each video, resulting in a database of 2000 labelled images. The video databases were simply composed of the entire frames from the dedicated videos. With these two types of databases computed for each surgery, we were finally able to evaluate both aspects of our framework, i.e. detection of the different visual cues and the global recognition rate of the entire framework.

Pre-processing validation

The first aspect of our framework that was validated was the spatial down-sampling. It was validated by comparing results of the alternative method for classifying visual cues. We don't present these results here, but internal studies have shown that the decrease of size by a factor of four had no impact on the classification results. It can be explained by the fact that the Carl Zeiss® microscopes provide videos with high quality that can be partially decrease without altering image processing performances.

The preliminary step of pupil segmentation, only applied to cataract videos, didn't require any training stage. This step was therefore simply validated over the entire video database by testing each frame of each video. During this validation, a pupil was considered correctly segmented if and only if the circle found by the Hough transform precisely matched the pupil. A percentage was then obtained corresponding to the accuracy of the segmentation.

Feature selection study

The feature selection method (as described in subsection IV.4.c), used only for the alternative approach of subsection IV.4, was applied to select the most discriminant features between phases. In this study, we computed for each visual cue from both datasets, the percentage of image features selected by the hybrid feature selection method by merging them into three categories: colour, form, and texture. For each visual cue, we therefore obtained the percentage of colour, form and texture features that was selected by the algorithm for further supervised classification.

BVW optimization

Before validating all visual cues recognition, we first optimized the BVW approach for both the recognition of the lens aspect and the detection of instruments presence. The goal was here to find the best combination between key-point detectors and key-point descriptors, as well as the optimal number of words for creating the vocabulary. For key-point detector, we tested 4 keypoints detection methods: SIFT, SURF, Harris and STAR, presented in subsection V.3.c that all provide access to local image information. All of them provided a similar result, which is a sample of keypoints, though they differed radically in the methods used to obtain them and by the nature of the keypoints found. For key-points descriptor, we focused on SURF descriptors for computational reasons. Indeed, the vector space dimension was reduced by a half (from 128 to 64) when switching from SIFT to SURF descriptors. This optimization was performed under the same conditions than the detection of other visual cues.

Validation of visual cues recognition

The second aspect of our framework that was assessed was the recognition of all visual cues for both data-sets. Only one type of visual cue classifier was not validated through cross-validation studies: the detection of the 1.4mm knife with the Haar classifier. In that case, the training stage was performed using manually selected positive and negative images for better object training. 2000 negative images and 500 positive images were used.

The four visual cues from the pituitary dataset, as well as the five (excluding the recognition of the 1.4mm knife) from the cataract dataset were assessed through 10-fold cross-validation studies. For

each surgery, the image database was divided into 10 randomly selected subsets. The subsets were composed of 300 images for the pituitary dataset, and 200 images for the cataract surgery. Nine were used for training while the prediction was made on the 10^{th} subset. This procedure was repeated 10 times and results were averaged. Accuracies, specificities and sensitivities were computed. After evaluating each visual cue classifier, we validated their use compared to a classical image-based classifier, i.e. compared to feature extraction, selection and classification as performed by the alternative method.

Entire framework validation

Lastly, we evaluated the global extended framework, including visual cue recognition and the time series analysis with the same type of cross-validation studies. Particularly, we used a leave-one-out method for validating both datasets. This involves that at each step of the cross-validation process, one video was used for the test while the others (15 in the case of pituitary dataset and 19 in the case of cataract dataset) were used for the training stage. For this assessment, the criterion chosen was the Frequency Recognition Rate (FRR), defined as the percentage of frames correctly recognized over a video by the recognition framework. The final results of the DTW approach were therefore compared to the HMM classification. In addition to this computation, the confusion matrix was extracted, showing exactly where states were misclassified.

V.6. Results

We present here the results of the different validation studies that we performed.

Pre-processing validation

Taking all frames from each video (at 1 fps), the pupil was correctly extracted with an accuracy of 95% (**Table 12**). The worse video was very difficult to segment, with 78% of all frames correctly segmented. The best video, on the other hand, had almost its entire frame correctly segmented (99%).

Table 12 - Mean, minimum and maximum accuracy of the segmentation of the pupil over the entire video database.

	Accuracy (Std)	Minimum	Maximum
Detection (%)	95(6)	78	99

Feature selection study

In **Figure 51**, we can see that the type of features selected for binary cues extraction really depends on the type of dataset. For instance, colour was the main category of features selected for the recognition of the microscope view for the pituitary dataset. Not surprisingly, it was not involved in the detection of the column of nose or nose retractor, where form features were more significant. For the cataract dataset, colour was the most important feature category selected for all binary cues.

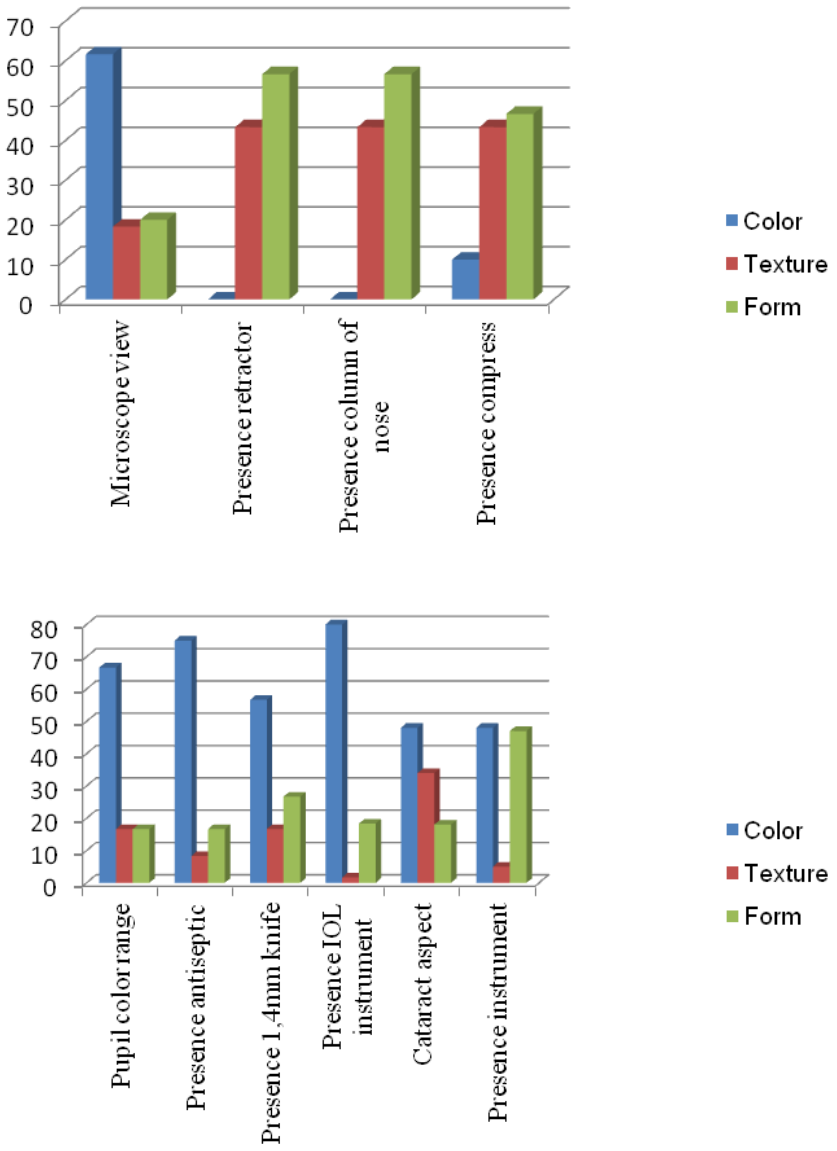


Figure 51 - Type of features (colour, texture or form) selected with the hybrid selection method for each binary cue. Below: Pituitary dataset. Above: cataract dataset.

BVW optimization

Figure 52 shows the BVW study for choosing the best parameters for both the detection of instrument presence and the texture-oriented classifier respectively. Surprisingly, for both figures, the number of visual words did not appear to be a major parameter to be enhanced. Indeed, the accuracy didn't vary significantly from 1 to 60 visual words, and this result was true for the 4 key-point detectors and the two BVW studies. For both studies, the best accuracy was still obtained for a number of visual words equal to 12. On the contrary, the influence of key-point detectors was significant. For the detection of instruments presence (**Figure 52**, above), the SURF keypoint detector showed best recognition

accuracies (with 12 visual words: ~86%), the SURF key-points detector shows best results, whereas for the detection of the lens aspect (**Figure 52**, below), the SIFT key-points detector outperformed other algorithms (with 12 visual words: ~83%).

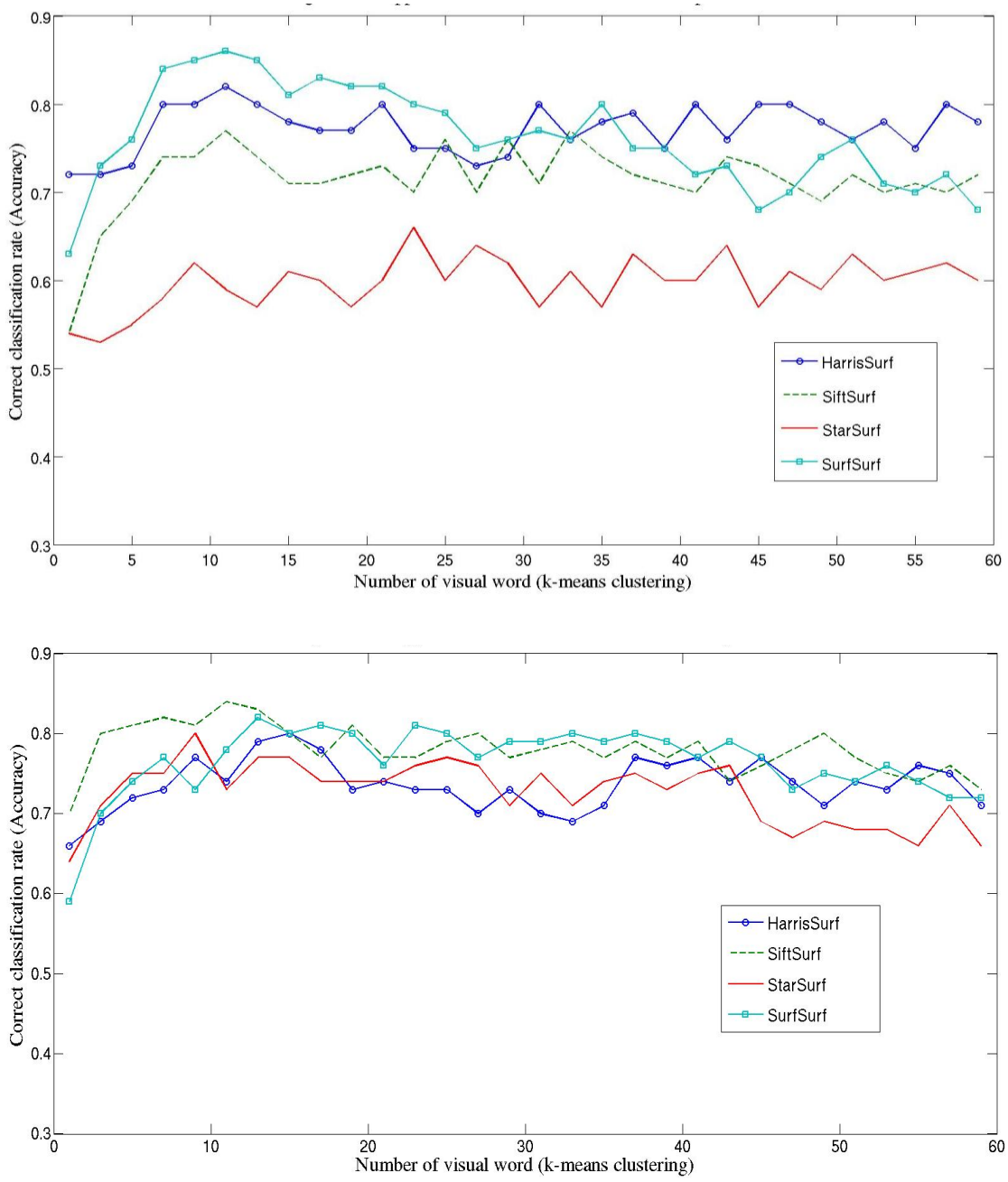


Figure 52 - BVW validation studies comparison of accuracies with different number of visual words and different keypoints detectors and descriptors. Above: detection of the instruments presence. Below: recognition of the lens aspect.

Validation of visual cues recognition

The results of the cross-validation study for the recognition of all visual cues (**Table 13**) showed that very good detection accuracies were obtained for both datasets using specific image-based classifiers, which often outperformed the classical classifier.

In particular, for the pituitary dataset, the detection of the zoom of the microscope as well as the detection of the presence of the column of nose showed reasonable results (88.9% and 94.8% respectively) with the traditional approach. On the contrary, the Haar classifier was not adapted to the detection of the nose retractors (65.2%), where the classical approach seems to be well sufficient to have a high detection rate (89.4). Finally, the detection of the compress using only colour histogram had sensibly the same accuracy than using the classical approach.

For the cataract dataset, all specific image-based classifier outperformed the classical approach. The best recognition was obtained for the presence of the 1.4mm Knife with the Haar classifier, achieving a recognition rate of 96.7%, whereas the lower rate was obtained for the recognition of the instrument presence (84.1%). The detection of the lens using a BVW approach on the pupil (the aspect of the lens) had not a very high accuracy (87.2%). Colour histogram approaches showed good results (96.2% for the pupil colour range detection and 96.1% for the antiseptic detection), whereas the IOL instrument had also a good recognition rate of 94.6%, even detected with the classical classifier.

Table 13 - Mean accuracy (standard deviation) for the recognition of the binary visual cues, using specific image-based classifier and using a classical approach. Above: Pituitary dataset visual cues. Below: Cataract dataset visual cues.

	Global-zoom view	Presence nose retractors	Presence column of nose	Presence compress
Specific image-based classifier (%)	88.9 (2.2)	65.2 (8.4)	94.8 (1.3)	87.5 (2.4)
Classical approach (%)	X	89.4 (1.1)	X	88.3 (1.6)

	Pupil colour range	Presence antiseptic	Presence Knife	Presence IOL instrument	Lens aspect	Presence instrument
Specific image-based classifier (%)	96.2 (3.6)	96.1 (0.7)	96.7 (3.4)	94.6 (1.1)	87.2 (5.4)	84.1 (8.6)
Classical approach (%)	94.1 (4.6)	95.6 (0.4)	88.5 (4.3)	X	54.1 (3.6)	58.7 (6.1)

Entire framework validation

The entire validation study (**Table 14**) showed that the pituitary dataset obtained a lower recognition rate than the cataract dataset using both the HMM or the DTW approaches, even with less phases to be detected and a highest duration time. From an algorithmic point of view, the time series study showed better results using the DTW approach than with HMM classification. Taking example on the cataract dataset, with HMM a mean FRR of 91.4% was obtained, whereas the DTW approach showed a mean FRR of 94.4%. Moreover, the results of the pituitary dataset were highly scattered (resulting in a high standard deviation), whereas results on the cataract dataset were more homogenous.

Other studies on this framework have shown that the maximum detection rate for both datasets using the DTW, obtained for the same video, was higher than 99%, whereas the lowest detection rate was also obtained on the same video and was lower than 75%.

Table 14 - Mean FRR of both datasets using the HMM and the DTW approaches.

	HMM	DTW
Pituitary surgeries	90.2 (6.4)	92.7 (4.2)
Cataract surgeries	91.4 (5.3)	94.4 (3.1)

A recognized sequence compared to the true sequence is shown in **Figure 53** for both datasets. In this particular example, each state was correctly classified with a maximum delay of 40s for the pituitary dataset and 10s for the cataract dataset.

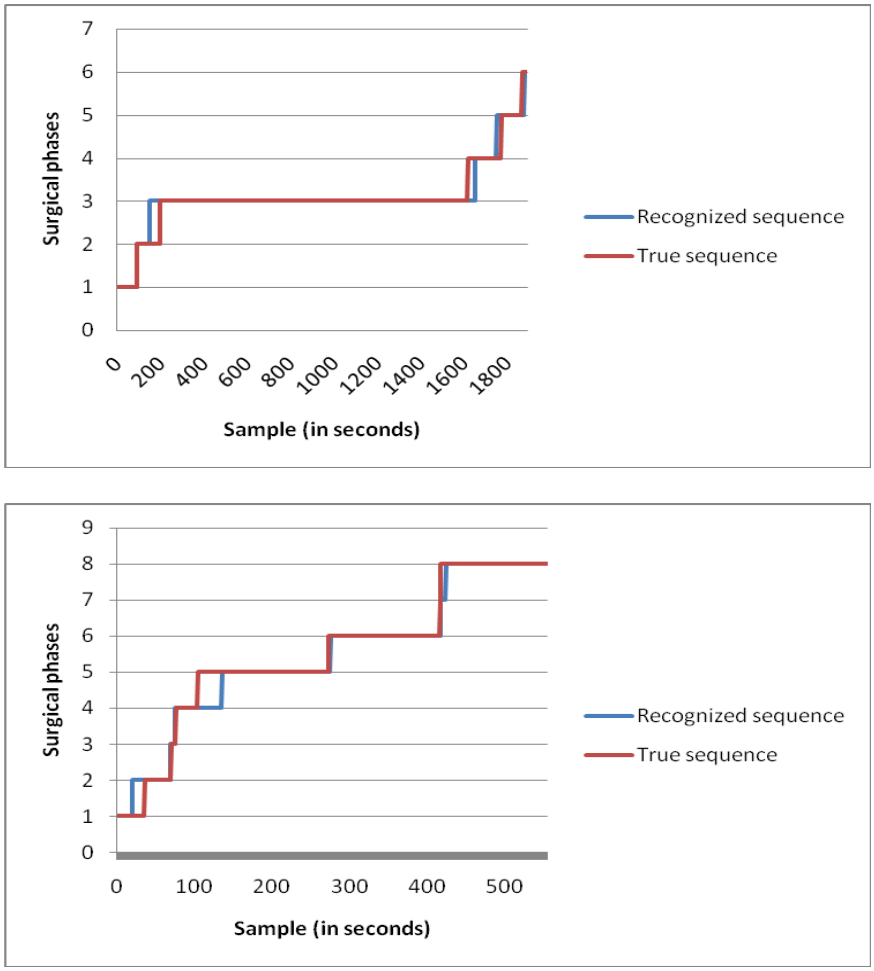


Figure 53 - Phase recognition of a video made by the HMM compared with the ground truth. Above: pituitary surgeries. Below: cataract surgeries.

From **Table 15** (above), we see that state n°3 contained the largest number of frames, and confusion was always between neighbouring states due to the sequentiality introduced by the HMM. The most significant error was for state n°5, where detection was approximately 75%. The highest accuracy (excluding the first and the last state) was for state n°4, where detection reached 95%. **Table 15** (below), shows similar results, but with more homogeneous phases. On this matrix, we see that errors can occur between non-consecutive phases, but that the majority of errors still occur between consecutive ones.

Table 15 - Confusion matrix for surgical phase detection with the HMM method. Rows indicate the surgical steps recognised and columns the ground truth. Above: pituitary dataset. Below: cataract dataset.

	1	**2**	**3**	**4**	**5**	**6**
1	5.68	0.97	0	0	0	0
2	0	4.68	4.09	0	0	0
3	0	0	72.99	0.2	0	0
4	0	0	0.45	3.04	0.07	0
5	0	0	0	0.04	3.31	0
6	0	0	0	0	0.99	3.49

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	4.1	0.3	0.1	0.7	0	0	0	0
2	0	5.6	0.1	0	0	0	0	0
3	0	0.2	0.5	0	0	0	0	0
4	0.3	1.2	0.4	6.4	0.1	0	0	0
5	0	0	0	4.8	22.1	3.1	0.1	0.3
6	0	0	0	1.1	1.6	22.4	0.3	0.2
7	0	0	0	0	0	0.1	0.5	0.2
8	0	0	0	0	0	0	0.1	23.1

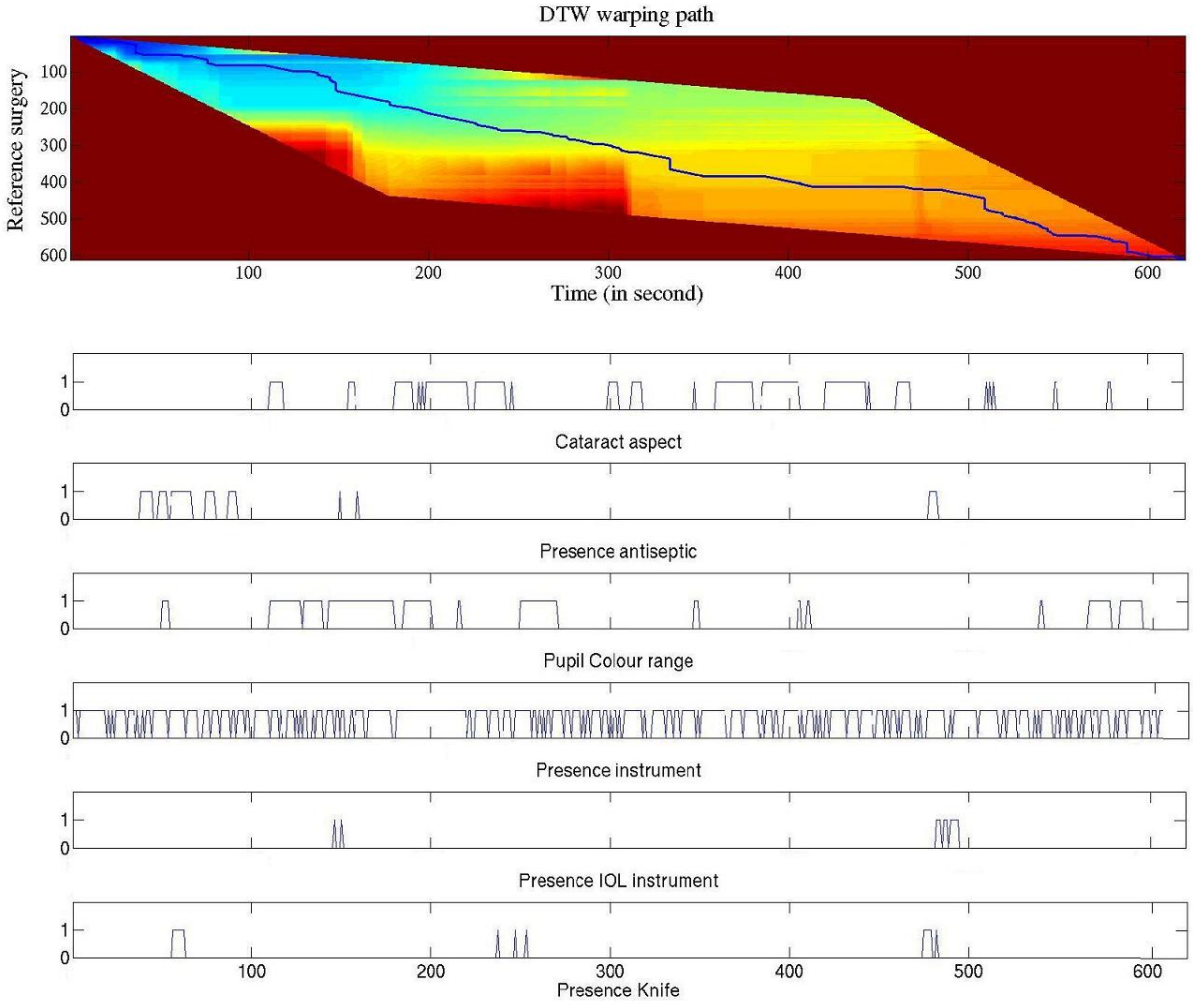


Figure 54 - Distance map of two surgeries and dedicated warping path using the Itakura constraint (above), and visual cues detected by the system (below).

Figure 54 shows an example of video recognized by the system with the warping path from the DTW approach and also with the different visual cues detected. The presence of instrument was not surprisingly the most frequent visual cue that is detected, whereas the presence of the IOL instrument or even the presence of the antiseptic were not often detected.

V.7. Discussion

In this chapter, we proposed a recognition system based on application-dependant image-based classifiers and time series analysis, using either an HMM or DTW algorithm. Using this framework, we are now able to recognize the major surgical phases of every new procedure. Compared to the previous chapter, we introduced on one hand local features that allowed an accurate definition of the video frames, and on the other hand the sequential aspect that allowed a better modelling of the phases of a surgery. This combined approach allowed a high degree of automatic recognition system accuracy. We have validated this framework with pituitary and cataract surgeries, where the sequences of surgical phases were recognized achieving recognition rate of around 92% for the pituitary dataset and 94% for the cataract dataset.

V.7.a. Content-based image classification

Our method for automatic surgical phase recognition addresses the well-known issue of the semantic gap, in which low-level visual features cannot correctly represent the high-level semantic content of images. Good classification requires an understanding of the important semantic categories that humans use for image classification and the extraction of meaningful image features that can correctly represent and discriminate these semantic categories. Moreover, required manual annotation for learning is time-consuming, and varies by user. Here, image annotation was a risky process, especially for binary cues definition. For this task, experts were asked to define the best combination of binary visual cues that efficiently differentiates surgical phases. For both datasets, the definition of binary cues was not exhaustive, and many other combinations may have provided the same phase recognition results. The multiplicity of solutions, along with the semantic gap issue, was the reasons why defining binary cues and assigning these labels to surgical microscope images remains a challenging task. On the other hand, the annotation of surgical phases was quite simple and quick. Manually capturing the semantic meaning of an image is much easier because it is close to human knowledge. Furthermore, both tested surgical procedures were highly reproducible and transitions between phases were distinctly defined, making the phase annotation process robust.

V.7.b. Pre-processing adaptation

The detection framework is based on the recognition of visual cues within the surgical scene. The recognition of such information allows the modelling of the surgical procedure and final surgical phase recognition. Due to the varying facilities between surgical departments, the numbers of phases, as well as the colours, tools and shapes could differ. Consequently, considering that surgical environments are different in each hospital, one recognition system should be tuned for each department. The adopted solution was to create a framework using specific image-based sub-systems in order to be as generic as possible, and to provide as many tools as possible for being exhaustive. This way, our method addresses the issue of system adaptability.

Even though the framework was created to be adaptable, each kind of surgical environment has its own particularities and characteristics. This is why preliminary pre-processing steps may be mandatory in order to tune the recognition system according to the type of surgery. For instance, in the case of low-resolution video images, the purpose would be to improve image quality for further processing. In the context of cataract surgery, the microscope field of view is precisely delineated, thus enabling the use of a preliminary step of segmentation to restrict the search for a specific visual cue within a ROI defined by the pupil outlines. This is the only step that is specific to cataract surgery. For the adaptation to other surgical procedures, this segmentation step could be either adapted or even removed. Taking as example neurosurgical procedures and specifically pituitary surgeries, segmentation would not be necessary as the field-of-view is already zoomed and adapted for image-based analysis and visual cues extraction. Pre-processing steps for image quality enhancement would also not be required because of the high-resolution of neurosurgical microscopes, neither intensity corrections nor specular reflection removal. This example would be true for this specific type of easy and reproducible surgical procedures. However, other adaptations could be conducted. Dealing with more complex surgeries would involve further researches on the pre-processing step, on the segmentation of surgical tools before their categorizations and possibly on the definition of other sub-systems for the detection of visual cues. We proposed in this chapter a complete framework that we tested on two surgical procedures, but the evolution to other surgical procedures should be experimented.

Once the framework has been tuned a dedicated surgical procedure, its use is fully automatic and will work with any microscope video of this type of surgery in its environment. Similarly, other variability factors may affect recognition, such as the manner in which surgeons operate. With this system, a training stage is necessary for each surgical department, assuming that the surgeons within the department use identical materials and follow the same sequence of phases during the procedure, making image features invariant to task distortion.

V.7.c. Pupil segmentation

Using an adapted method composed of image-based analysis, the segmentation of the pupil provides highly accurate results. For 95% of the frames, the ROI correctly contained the entire pupil. Moreover, to avoid distorting any further detection that could be done within the pupil, we decided to define a constant diameter value. Thus, each time a ROI was detected, the centre was kept and the circumference value was reset to the default value. Due to its high accuracy over the entire video database, it allows all potential colour-associated noise to be removed from around the pupil for further recognition. The very low accuracy obtained for one video can be explained by the presence of the retractors, rendering the field of view very narrow. Automatic segmentation turns out to be difficult when the retractors, or even the surgical instruments, around the eye, occupy too much space within the field of view.

Incomplete circle outlines in the image mask may occur, leading to Hough circle detection failure. For this 5% failure detection, where pupils were not correctly segmented, an alternative simple method could be used. This method could also be based on the binary mask created in **Figure 42**. The idea is to go through the mask and search for the circular zone of a predefined diameter that contains the most number of white (or black) pixels. It is an iterative search that identifies the most probable circular zone (**Figure 55**).

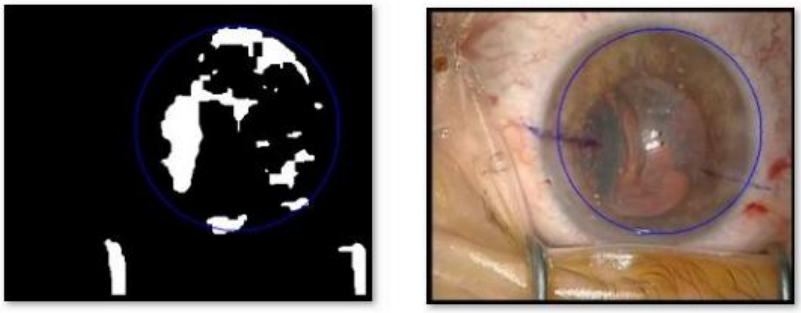


Figure 55 - Alternative method to the Hough transform.

Unfortunately, due to the difficulty of detecting errors in the recognition process using the Hough transform, this method was not implemented within the global framework.

As another drawback, our approach, which always returns a ROI, was not always perfectly centred on the middle of the pupil. We can explain this issue by the fact that the pupil was not always completely inside the microscope's field of view. Sometimes the pupil outlines were too distorted due to surgical tools or the surgery itself. Sometimes the retractor was as wide as the pupil and sometimes the surgeon's fingers were in the field of view. In that case, it was difficult to extract the exact position of the pupil and its outlines and to adjust an intensity threshold accordingly. If the surgical microscope had a permanent position, or if we could precisely estimate the position of the pupil in each image, it would be possible to automatically adjust a threshold for the segmentation.

V.7.d. Application-dependant visual cues

The purpose of this step of our procedure was to extract relevant binary cues from the microscope images that can differentiate surgical phases. In other words, the surgeon was asked to define visual information that was linked to each specific phase. It was requested that these binary cues be easily identifiable through image features analysis, and that only two values (binary signal) be taken. A few pieces of information that are relevant for the detection of SPMs were removed and replaced because they were not detectable with a standard image-based analysis.

Before the visual cue recognition training stage, the user will need to choose the visual cues and the associated image-based recognition classifier. In image classification problems, users usually do not think in terms of low-level features, resulting in poor recognition of the high-level semantic content of the images. Here, during the stage of visual cue definition, the colour, texture and shape behaviour of the visual cues are often intuitively known, allowing the most effective classifiers to be chosen. When visual cue is unknown or undocumented, the solution proposed is to choose the generic approach, integrating a large number of image features. This approach, combining global spatial features and SVM, may therefore be adapted to the recognition of any type of cue. The feature selection step allows the user to select discriminatory features and remove unsuitable ones, which is the intended objective. To improve recognition, however, the three other specific classifiers seem to be well-adapted when the behaviour of the visual cue is well perceived.

Generally, the main drawback of a global colour histogram representation is that information concerning object shape, and texture is discarded. In our case, however, it was only used for colour-

related visual cue detection. Similarly, the main drawback of shape-based approaches is the lack of well-defined outlines. The Haar classifier was used in our framework for specific objects only, e.g. the knife, which is significantly different from all others instruments used in cataract surgery. The use of this approach to categorize other instruments, such as the cannula, was tested, but gave very poor results due to the narrow field of view and the difficulty in discriminating that specific instrument from the others. For this reason, we chose to use a second approach for object recognition, allowing the system to gain information concerning object presence, without categorizing it. This type of information was still relevant for phase detection and allowed complete image signatures to be achieved using other information with a different level of granularity. The use of a BVW approach combined with local descriptors was also validated. Local descriptor comparisons enabled selection of the most appropriate features, and application with the recognition of the global aspect of the lens gave very promising results.

With the exception of the Haar classifier, the three other classifiers are all based on a training image database. The power of discrimination of the image database is thus vital. We can easily imagine that accuracy may decrease sharply if the images do not efficiently represent all phases or all scene possibilities within the phases. Additionally, the training stage is time-consuming and requires human efforts. In our particular case, the best method, used here in our validation studies, was to annotate surgical videos before randomising the initial sample. The randomisation process is thus no longer performed on all frames, but on each video independently, extracting the same number of frames per video.

V.7.e. Time series analysis

Combined with state-of-the-art computer vision techniques, time series analysis showed very good performance, opening the way for further promising work on high-level task recognition in surgery. Without this step of time series analysis, results are less accurate (~80%). This can be explained because some visual cues don't appear during one particular phase only, and the information of sequentiality is needed. For instance, the knife always appears twice during cataract surgery: once during phase n°4 (principal corneal incision), and once during phase n°10 (expansion of the principal incision). All other visual cues are not present during these two phases. The discrimination of both phases appears to be possible with an information of time only that the HMM or the DTW can bring.

In particular, DTW captures the sequence of surgical phases and is well-adapted to this type of detection. The cost function between 2 surgical procedures with the same sequence of phases, but with phase time differences, will be very low. The advantage is that it can accurately synchronize two surgical procedures by maximally reducing time differences. The main limitation concerning the use of DTW, however, is the phase sequence differences that may appear between two surgeries. The warping path would not correctly synchronize the phases and errors would occur. In the context of cataract surgery, the procedure is standardized and reproducible, justifying the very good results of the recognition. But we can imagine that for surgeries that are not completely standardized, DTW would not be adapted. In this case, HMM could be used by adding bridges between the different states of the model (the transition matrix should be adapted in that case), allowing the sequence to be resumed and to perform the same phase multiple times. The state machine would not be a left-right structure but would include more complex possibilities with many bridges between states. As a drawback, the complexity of such HMMs could completely affect the recognition accuracy. For each surgical procedure, the HMM structure should be created by minimizing the possibilities of transitions from

states to states not to affect the classification phase. In the particular case of cataract surgery, the results showed that the DTW algorithm was, not surprisingly, quite better than HMM.

Another limitation appears when an adverse-event occurs in the OR, which never happened in any video of the training sample. In such cases, image signatures are not linked to any phase and it could affect the recognition process. Using HMM, it could affect the transition probabilities and therefore the entire recognition system. Using DTW algorithm, it could modify the warping path and therefore alter the detection of surgical phases. One solution to this issue, in the case of HMM classification, would be to detect such images (containing features that are very different from others, and therefore easily detectable) and to create a specific state for unknown images in the HMM, as done in Padoy et al. (2008). This state would be connected to all other states and the path would cross this state every time image features are too different from the training sample.

V.7.f. Temporal features

We tested both approaches presented in subsection V.4 on the cataract surgery images.

For experimenting the spatio-temporal key-points detection method, a set of parameters were empirically chosen: $\sigma_t = \sigma_s = 2$, $k = 0.04$. At each time step, the current video frame was analyzed with the previous and next frame (after the pre-processing step, i.e. the video at 1Hz). The analysis was therefore performed over a period of 3 seconds, allowing surgical tools to have significant displacements in the video. The key-points detection threshold was set to 6000. An example of spatio-temporal features on one video frame is shown on **Figure 56**.

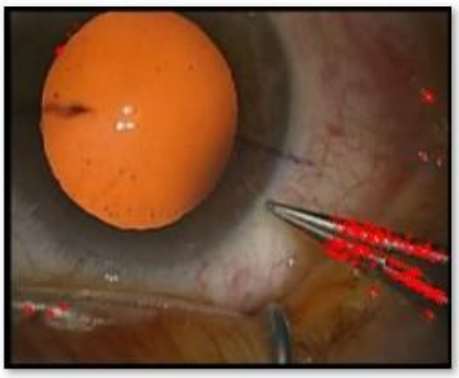


Figure 56 - Illustration of spatio-temporal features obtained with our parameters.

Even if these key-points were intuitively accurately extracting movement information, such as the movement of the colibri tweezer in **Figure 56**, this information was difficult to integrate into our framework. During the visual cues extraction step or during the time-series modelling step, no optimal solution was found to improve the recognition rate thanks to this temporal feature information. One possibility would have been to use the STIP key-points as a preliminary step for detecting the different surgical tools before launching the BVW approach. The problem of STIP key-points is that it also detects the background movement, and the differentiation between pixels belonging to a surgical tool or pixel belonging to the background was hard to identify. We therefore decided not to integrate this method into the framework.

For experimenting the optical flow method, we choose the Farneback algorithm (Farneback, 2001), which is very close to the Horn-Schunck algorithm. The analysis was also performed over a period of 3 seconds. Using this method, a displacement vector was computed for each pixel of the frame (**Figure 57**).

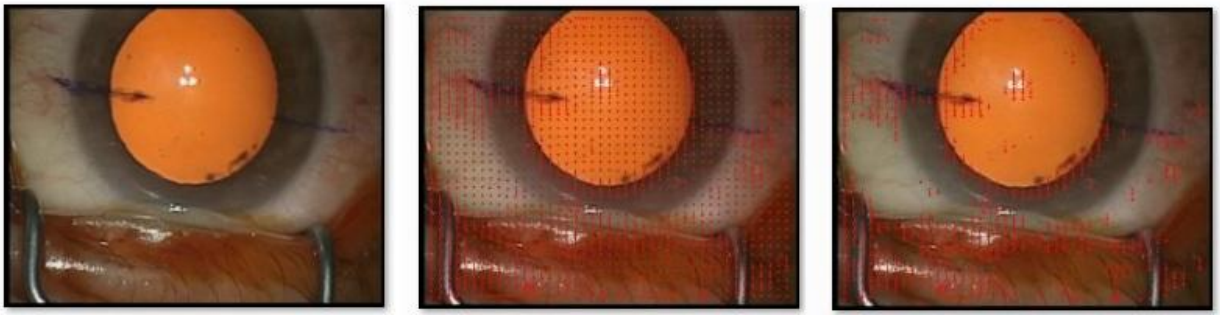


Figure 57 - Illustration of the optical flow method at time t, t+1 and t+2 (from left to right).

After the computation of displacements vectors for a set of video frames, one way to analyse the results is to extract the primary movements. First, null displacement vectors as well as light ones were removed. The computational benefit was non-negligible. Thresholds were fixed to $\delta x = 5$ and $\delta y = 5$. Then, a clustering (i.e. non-supervised classification) was performed on remaining displacement vectors by integrating a spatial component (**Figure 58**). Displacement vectors were transformed into one value corresponding to the angle, and forces of displacement were removed. Instead of having 5 features (position of the pixel, displacement vectors and force), the clustering was performed over 3 features in order to take into account a strong spatial component. K-means technique was used and the number of cluster was incremented until intra-classes variability was superior to a pre-defined threshold. Results showed the extraction of between 2 and 5 classes per image, corresponding to the different objects of the images and to the background displacement. The last step consisted of averaging displacement vectors of each class in order to have one primary movement vector per object.

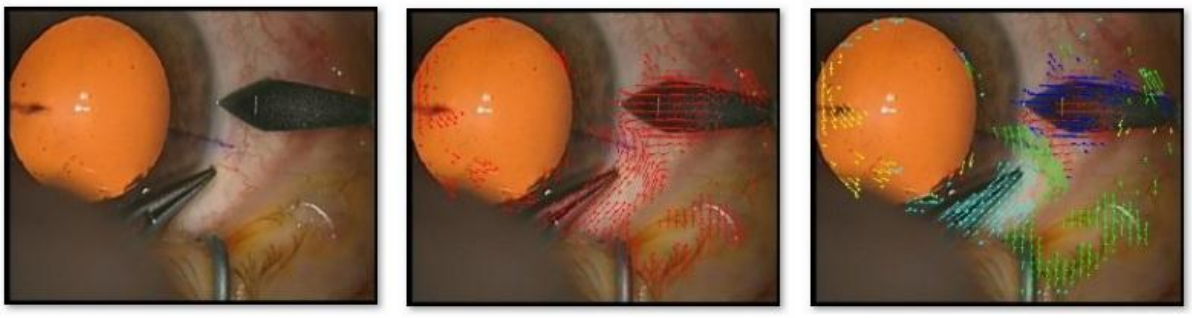


Figure 58 - Illustration of the clustering applied on displacement vectors. On the right image, the dark-blue class corresponds to the displacement of the 1.4mm knife, the light-blue class to the colibri tweezers, the green class to the global background displacement and the yellow and red ones to other background elements.

The idea of having one displacement vector per object is to associate this result to the recognition of the surgical tools. To make the link between both analyses, surgical tools are associated with

displacement vectors that are spatially close. Moreover, the general background displacement could be extracted and its value should be subtracted to each displacement vectors of objects for computing the absolute displacement vectors of surgical tools. Similarly to the detection of STIP key-points, results of this approach, even very interesting, presented some errors and turned out to be less efficient than the segmentation into 2 ROIs of the surgical tools. We also didn't use it into the recognition framework.

V.7.g. From high-level tasks to low-level tasks recognition

We proposed in this Chapter a recognition system based on application-dependant image-based classifiers and time series analysis, using either an HMM or DTW algorithm. Using this framework, we are now able to recognize the major high-level tasks of every new surgery that would have been previously adapted to the framework. After this successful step, we decided to experiment the detection of tasks with a lower granularity level. Even if the recognition of high-level tasks would be interesting for many clinical applications (see the general discussion of subsection VII.3), addressing the recognition of surgical tasks at lower granularity levels from microscope videos only is challenging and of great interest for SPM methodology. Image-based analysis may not be sufficient for this type of detection, but we will see in the next Chapter that the addition of top-down reasoning to the traditional bottom-up analysis is a promising way of addressing this challenge.

References

- ✓ Agrawal M and Konolige K. CenSurE: Center surround extremas for realtime feature detection and matching. European Conf Comput Vision, ECCV. 2008; 5305: 102-15.
- ✓ André B, Vercauteren T, Perchant A, Buchner AM. Endomicroscopic image retrieval and classification using invariant visual features. Proc. ISBI. 2009. 346-9.
- ✓ Bay H, Tuytelaars T, Van Gool Luc. SURF: Speeded Up Robust Features. European Conf Comput Vision, ECCV.2006.
- ✓ Beauchemin S and Barron JL. The computation of optical flow. ACM..1995.
- ✓ Bhuyan MK, Ghosh D, Bora PK. State selection via DTW for finite state based gesture recognition. National Conference on Communications. 2006: 265-8.
- ✓ Brunelli R. Template Matching Techniques in Computer Vision: Theory and Practice. Wiley. 2009.
- ✓ Box G and Jenkins G. Time series analysis: Forecasting and control. Holden-Day. 1970.
- ✓ Canny J. A Computational Approach To Edge Detection. IEEE Trans Pattern Analysis Mach Intell. 1986; 8(6):679-98.
- ✓ Cuntoor NP, Yegnanarayana B, Chellappa R. Activity modeling using event probability sequences. IEEE Trans. Image Processing.2008.
- ✓ Dame A and Marchand E. Optimal detection and tracking of feature points using mutual information. IEEE Int Conf on Image Processing - ICIP. 2009.
- ✓ Farneback G. Very high accuracy velocity estimation using orientation tensors, parametric motion models, and simultaneous segmentation of the motion field. IEEE Int Conf Computer Vision. 2001; 1: 171-7.
- ✓ Fleet DJ, Weiss Y. Optical Flow Estimation. Handbook of Mathematical Models in Computer Vision. Springer. 2006.
- ✓ Fox C. An introduction to the calculus of variations. Courier Dover Publications. 1987.
- ✓ Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Proc European Conf on Computational Learning Theory. 1995.
- ✓ Golub GH and Van Loan CF. Matrix Computations. Johns Hopkins University Press Baltimore, MD, USA. 1996.
- ✓ Harris C, Stephens M. A combined corner and edge detector. Alvey vision conference. 1988.
- ✓ Horn BKP and Schunck BG. Determining optical flow. Artificial intelligence. 1981; 17: 185-203.
- ✓ Hough VC. Machine Analysis of Bubble Chamber Pictures. Proc Int Conf High Energy Accelerators and Instrumentation. 1959.
- ✓ Kalman RE. A New Approach to Linear Filtering and Prediction Problems. Transaction of the ASME—J Basic Engineering. 1960; 35-45.
- ✓ Ke Y and Sukthankar R. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. Computer Vision and Pattern Recognition. 2004.
- ✓ Keogh EJ, Pazzani MJ. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. Prediction of the future: AI approaches to time-series problems. 1998; 44-51.
- ✓ Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Int Conf Machine Learning. 2001.
- ✓ Laptev I, Lindeberg T. Local descriptors for spatio-temporal recognition. Spatial coherence for visual motion analysis. 2006. Springer
- ✓ Lowe DG. Object recognition from scale-invariant features. ICCV. 1999; 2: 1150-7.

- ✓ Lowe DG. Distinctive image features from scale-invariant keypoints. International journal of computer vision. Springer. 2004.
- ✓ Lucas BD and Kanade T. An iterative image registration technique with an application to stereo vision. Proc Imaging Understanding Workshop. 1981; 121-30.
- ✓ Mikolajczyk K and Schmid C. A performance evaluation of local descriptors. IEEE Trans Pattern Analysis Machine Intell. 2005; 10(27): 1615-30.
- ✓ McCallum A, Freitag, D, Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation. Proc. ICML. 2000; 591-8.
- ✓ Niennattrakkul V and Ratanamahatana CA. Learning DTW global constraint for time series classification. Artificial Intelligence papers. 1999.
- ✓ Padoy N, Blum T, Feuner H, Berger MO, Navab N. On-line recognition of surgical activity for monitoring in the operating room. Proc Conf Innovative Applications of Artificial Intelligence. 2008.
- ✓ Rabiner LR. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc IEEE. 1989; 77(2).
- ✓ Roberts LG. Machine Perception of Three-dimensional solids. Massachusetts Institute of Technology, Lincol Laboratory. 1965.
- ✓ Senthilkumaran N and Rajesh R. Edge Detection Techniques for Image Segmentation - A Survey. Proc Int Conf on Managing Next Generation Software Applications (MNGSA). 2008; 749-60.
- ✓ Shubin MA. Laplace Operators. In Hazewinkel, Michiel, Encyclopedia of Mathematics. Springer. 2001.
- ✓ Smeulders A, Worrin M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Pattern Analysis Machine Intelligence. 2000; 22(12): 1349-80.
- ✓ Viola P. and Jones, M. Rapid real-time face detection. IJCV. 2004; 137-54.
- ✓ Viterbi A. Errors bounds for convolutional codes. IEEE TIT. 1967; 13(2): 260-9.
- ✓ Wang K and Gasser T. Alignment of curves by dynamic time warping. Annals of Statistics. 1997; 25(3): 1251-76.
- ✓ Xiang T, Gong S. Beyond tracking: modelling activity and understanding behavior. Int J Comp Vision. 2006; 67(1): 21-51.

Chapter VI. Surgical activities detection *knowledge-based approach*

We have seen in Chapter II that extensive work has already been performed on the creation of SPMs, though a few studies have focused on the automatic recognition of low-level surgical tasks, i.e. activities from sensor data (Kragic and Hager, 2003; Speidel et al., 2008; Agarwal et al., 2007; Houliston et al., 2011; Miyawaki et al., 2005; Yoshimitsu et al., 2010). In such cases, a hierarchical decomposition of the procedure is always provided in order to establish a link between the different granularity levels. Only a few of these studies attempted to automatically recognize the activities from low-level data, and none of them used videos. In this last Chapter, we show how we further extended our approach by going one level down on the granularity axis toward the detection of surgical activities. This level of granularity is symbolized by the use of one surgical tool for one surgical activity performed on one anatomical structure, as seen in Chapter III. One statement on which we relied for this new study was the following. The sequential nature of the surgical phases that we detected in previous Chapters can serve as a temporal constraint for activity detection. Indeed, even if activities do not follow any strong sequential behaviour that inhibits the use of any time-series algorithm, the majority of activities occurs in one or two phases only, limiting the possible number of activities per phase. Based on this statement, and using a 2D graph formalized as a hierarchical decomposition of the surgery with phases and activities, surgical activity detection becomes feasible. In addition to the phase information, we adapted the surgical tool detection algorithm based on BVW approach and the pupil segmentation process that were both presented in the previous Chapter to this problem. Knowing the surgical phases before the activity recognition and instead of using one supervised classification algorithm for the entire video, one supervised classification was launched per phase. The addition of knowledge to the framework makes the issue of activity detection easier, combining a traditional bottom-up approach with top-down reasoning. The main novelty of this approach is that we tackle the problem of low-level task extraction using existing sensors in the OR. Following the results of Chapter V, we experimented our methods on cataract surgeries only. This type of surgery offers the possibility to be very standardized and allows the definition of surgical activities.

VI.1. Methods

For the automatic detection of activities, and according to the formalization introduced in subsection III.2.c where each activity is defined by a triple *<action, surgical tool, anatomical structure>*, the three components of an activity should be extracted for accurate recognition. Unfortunately, the actions, represented by verbs describing the movements of the surgeon's hands, are very hard to identify and classify. Studies on movement detection were conducted in subsection V.7.f, but without any satisfactory results. We therefore focused on the two other components. Then, a hierarchical

decomposition was proposed, making the link between high-level (surgical phases) and low-level (surgical activities) tasks in the context of cataract surgeries.

VI.1.a. Pre-processing

The same spatial downsampling than the previous chapter was performed on the cataract videos (i.e. decrease by a factor of 4), but the spatial downsampling was different. Indeed, the detection of lower activity level requires a more precise time description in order to capture all dynamic information. We therefore decided to downsample to 2fps.

VI.1.b. Surgical tools detection

Detecting and recognizing surgical tools in cataract surgery can be relatively complex using image-based analysis due to the similar shapes of tools, and to differences in orientation, scale, or illumination. The method we propose here is able to automatically detect and categorize tools, and is entirely based on the first work presented in subsection V.5.c Here, instead of defining only two classes for categorizing the ROIs as being a surgical tool or only background, we defined 7 classes including six for the surgical tools and one for the background class that doesn't contain any tools. Many of tools in cataract surgeries are highly similar, and one class cannot be created for each tool. Specifically, 4 tools are very similar and hard to distinguish: the irrigation cannula, the sauter canula, the aspiration cannula and the micro-spatula. These 4 tools were regrouped into one class. Finally, six classes were defined: class n°1: Irrigation cannula, Sauter cannula, Aspiration cannula, Micro-spatula; class n°2: 1.1 mm knife; class n°3: Methocel tool; class n°4: Wecker scissors; class n°5: colibri tweezers; class n°6: Chopper, and class n°7: background. Examples of these surgical tools are shown on
Figure 59.

Figure 59 - Examples of surgical tools used in cataract surgeries. Left to right: colibri tweezers, wecker scissors, 1.4mm knife, micro spatula, aspiration cannula and 1.1mm knife.

Each of the 7 classes was built using 100 representative images from the dataset that were manually chosen. Following results of the first study using BVW for the detection of instruments presence (V.6), we kept the SURF key-points detector, the SURF key-points descriptor and we also kept 12 visual words according to the results of **Figure 52**. Moreover, instead of using a SVM classifier that is well suitable for binary classification, we used a KNN algorithm (k=5) in order to obtain for each ROI its probability of containing each surgical tool. We finally obtained a percentage of belonging to each class, which appears to be more flexible than using a strong SVM classifier. For instance, some pairs of surgical tools have very similar shapes (e.g. micro the spatula and the aspiration cannula) and having a percentage gives much more information that having only its most likely class.

VI.1.c. Anatomical structures detection

The other useful information to be extracted is the anatomical structure on which the tool is used. As depth information is missing from microscope videos, 3 zones were identified on each image, based on the segmentation of the pupil presented in subsection V.5.b. The first zone corresponds to the pupil in the centre of the image; the second zone includes the iris around the pupil, whereas the third zone is the remaining part of the image. For segmenting the three zones, the first step consisted in the segmentation of the pupil previously proposed. On 100 images, the pupil and the iris were manually segmented, resulting in an average radius for the second zone corresponding to the iris. This second constant value was also applied to the circle centre for determining the second and the third zones (**Figure 60**). Using this segmentation, and knowing the exact shape and location of the two ROIs previously extracted, we could affect an occupation percentage of the three zones to each ROI.

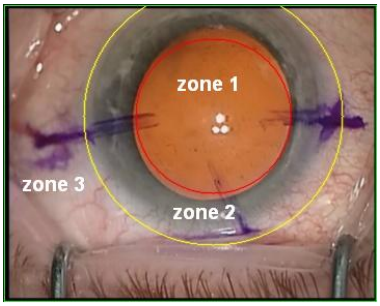


Figure 60 - Illustration of the three zones: zone 1: pupil, zone 2: iris, zone 3: rest of the image.

VI.1.d. Colour-based activity detection

Three activities remained undetectable through surgical tool detection only:
<*implant, IOL injector, anterior chamber*> (**Figure 61** – left)
<*swab, swab pagasling, conjunctiva & cornea*> (**Figure 61** – middle)
<*disinfect, betaisodona tool, conjunctiva & cornea*> (**Figure 61** – right)

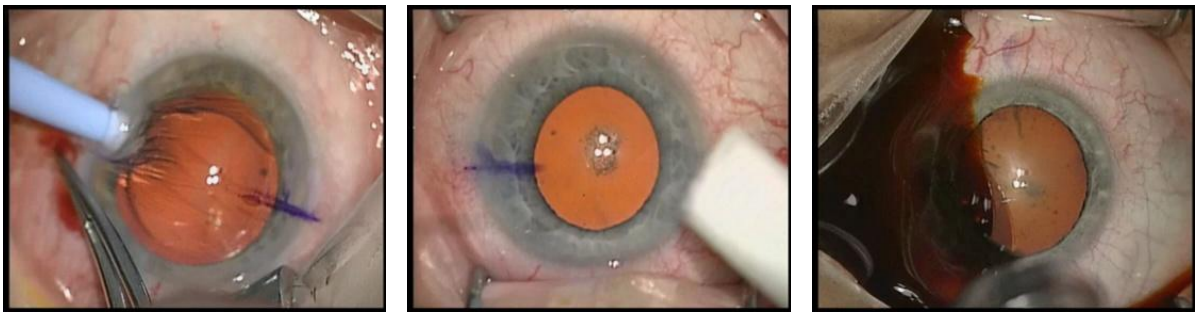


Figure 61 - Example of the three activities that are undetectable through surgical tool detection only.

The first activity is very difficult to detect with a BVW approach because the IOL injector is hard to segment due to its relative transparency on its tip. The second activity is the only one that is not composed of a surgical tool. It was included into the activity terminology because it appears many times during a cataract procedure and this gesture can be assimilated to a specific activity. Finally, the third one is also difficult to detect because the betaisodona tool may not appear into the microscope

field-of-view. Even if these 3 activities were not identifiable by their surgical tools, they can be detected through their particular colour. The IOL injector was recognizable by its blue colour, the swab by its white colour, and the betaisodona by its red colour. A colour histogram approach was therefore used for the recognition of these 3 visual cues and integrated into the final image signature.

VI.1.e. Knowledge-based supervised classification

Activity recognition was based upon the fact that most activities occur in only one or two phases, thus limiting the scope of activity possibilities per phase. Therefore, a 2D graph formalized as a hierarchical decomposition of the surgery with phases and activities was created. For each of the 8 phases of a cataract procedure, we associated its set of possible activities (
Figure 62). At this point in the analysis, information concerning surgical tools and their corresponding zones could be extracted from each image comprising a signature of 23 features: 2 ROIs * (7 surgical tools classes + 3 zones) + IOL injector detection + swab detection + Betaisodona detection. Instead of using one training stage for the entire video using all pairs of activities, knowing the surgical phases before activity recognition enabled us to launch one supervised classification per phase. Surgical phase recognition isn't fully accurate (accuracy~94%), but errors always occur between consecutive phases. In order to take this error into account, each supervised classification will contain pairs of activities of the on-going phase, along with pairs of activities from the previous and next phases. Including activities from adjacent phases will still permit to decrease the set of possible values. Instead of the 25 initial pairs of activities, each training was launched using between 6 and 13 classes (average=8 classes). The multiclass SVMs algorithm (Crammer and Singer, 2001) was chosen for the supervised classification.

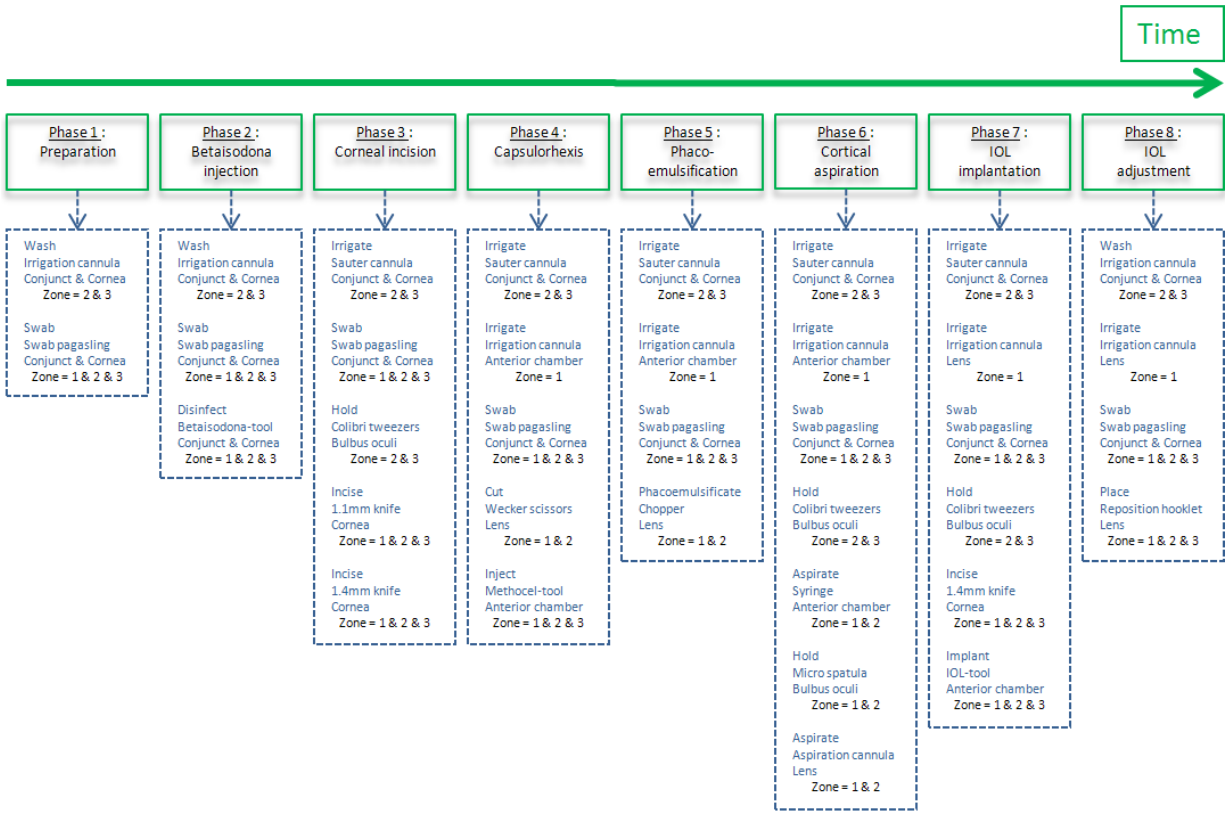


Figure 62 - Surgical phases and their corresponding possible activities.

VI.2. Results

The average recognition rate, taken from the entire system was not very high with a value for the frame-by-frame detection of **64.5 +/- 6.8%** (**Table 16**). The sensitivity (76.6%) is also better than the specificity (54.9%).

Table 16 - Mean FRR, specificity and sensitivity of the surgical activities.

	Mean FRR (std)	Specificity	Sensitivity
Accuracy (%)	64.5 (6.8)	54.9	76.3

We noticed, based on **Figure 63**, that there was a strong recognition disparity between pairs of activities. The two major pairs of activities were correctly identified, with detection rates greater than 70%. Other activities show good detection accuracies, such as the <*swab, swab-pagasling, conjunctiva & cornea*> or <*irrigate, irrigation-cannula, anterior-chamber*> activities, but these activities occurred during less than 10% of total surgery time. The detection rates of some activities, on the other hand, were very low and even close to zero for <*hold, colibri-tweezers, bulbus-oculi*>. We also noticed that three activities had occurrence percentages close to zero, whereas 7 activities appeared during less than 1% of total surgery time. The results of **Figure 63** illustrate the low sensitivity value that we previously obtained.

VI.3. Discussion

We proposed in this Chapter to tackle the problem of surgical activity recognition in the OR using existing sensors, by adding *top-down* reasoning to the traditional *bottom-up* approaches. Based on a hierarchical decomposition of the procedure containing the links between activities and phases, we used a multiclass SVMs algorithm to classify each frame of surgery videos. In addition to the phase information, image signatures were also composed of surgical tool information, including categorization of the tools and zone segmentation. We validated our methodology with the dataset of cataract surgeries. Eighteen activities were identified, containing up to 25 pairs of activities. A frame-by-frame recognition rate of 64% was achieved using leave-one-out cross-validation.

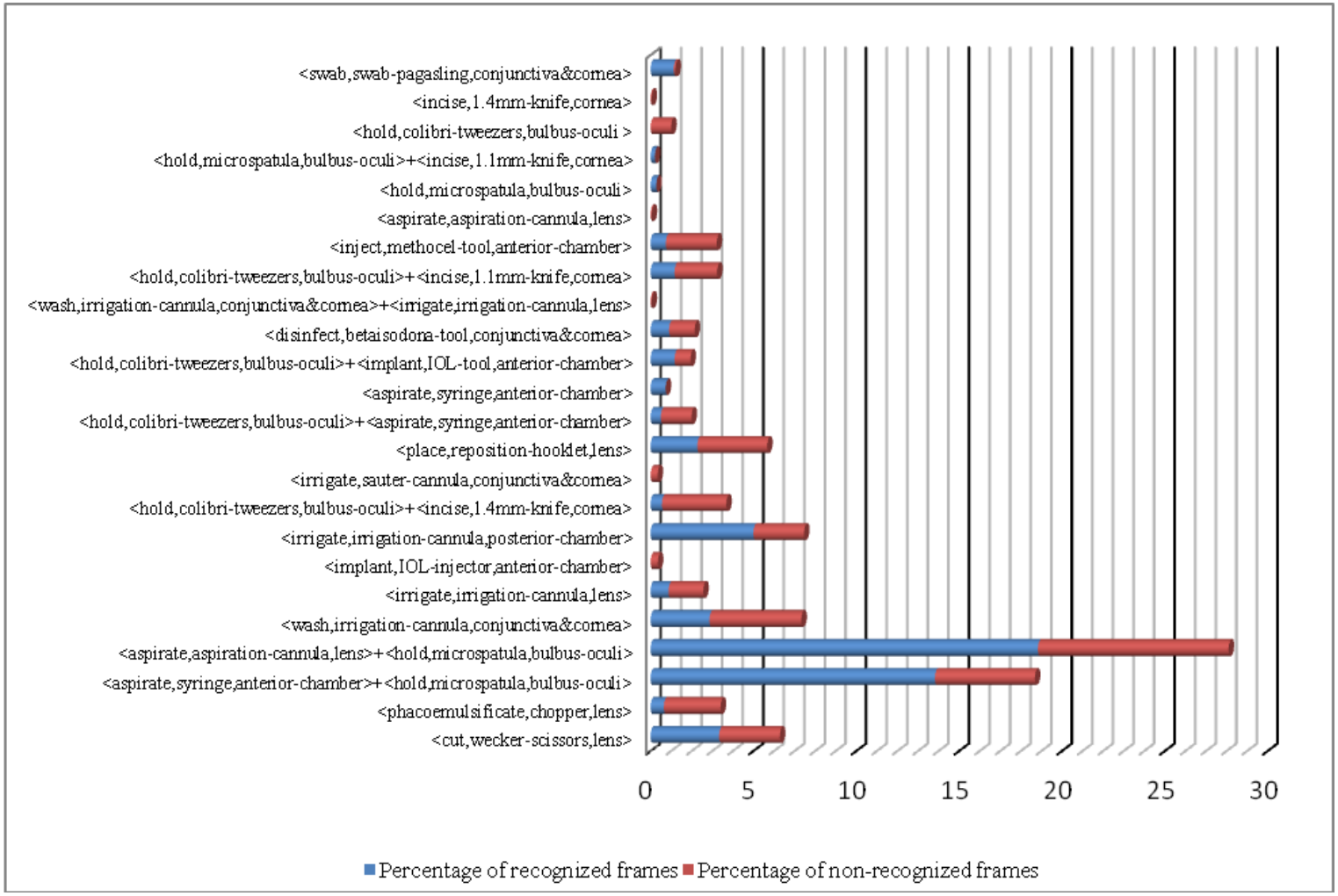


Figure 63 - Percentage of recognized and non-recognized frames for each possible pair of activities, normalized over the entire data-set (i.e. percentage of total surgery time).

VI.3.a. Action detection

The first component of an activity, the action, is impossible to identify using image-based analysis only. Spatio-temporal analysis (STIP points; Laptev and Lindeberg, 2006) and optical flow (Beauchemin and Barron, 1995) were tested for the extraction of trajectories, but no particular patterns were identified for each action. For instance, the 3 actions *irrigate*, *wash*, and *disinfect*, or the 4 actions *replace*, *aspirate*, *implant*, and *inject* have identical trajectory patterns and very close spatio-temporal features. Even between these groups, differentiation remained difficult. Moreover, the trajectory may differ according to the surgical tool used within the activity, rendering the extraction of patterns for each action difficult. Over the set of 6 actions, the only one that involved a unique pattern was the action *cut* belonging to the activity <*cut, wecker scissor, lens*>. In that particular case, the surgeon executes a circular trajectory in order to perform the capsulorhexis phase. In all other cases, surgical tools enter very quickly into the surgical scene with a linear trajectory, execute some small slow movements in the same area and go out of the scene with a linear trajectory. Another main issue for action recognition is that the microscope doesn't remain immobile, and the pupil is never in the same position within the field of view. Consequently, in addition to the detection of action patterns, detecting the background movement pattern (due to microscope displacement) would be required. It would finally be necessary to combine these two movement patterns to be sure to isolate an action pattern only. Because of these difficulties, efforts have focused on surgical tool and anatomical structure detection.

VI.3.b. Surgical tools detection

Using information obtained from the connected components, and hence from the surgical tools, is a logical option as each activity is linked to the use of a surgical tool. However, imperfections in surgical tool detection have a direct impact on the final supervised classification result. Tool detection is therefore the aspect of our work that needs the most significant improvement. This step, like zone segmentation, has been tuned for this type of surgery, and shows reasonable results (84.1%). Some tools are easier to segment by virtue of their bigger size (e.g. *1.1 mm knife*), or because of more important color gradients (*e.g. irrigation cannula*). A drawback is that connected components obtained during the first stage of the method did not always contain whole surgical tool. This incomplete detection induced lower recognition rates. Moreover, it was quite difficult to build a complete background class (i.e. class "not a tool"). Random images were chosen but may not correctly represent the field of possibilities of the background. Another limitation was the high number of class 1 tools. The highly similar shapes of class 1 tools prevented the creation of one class per tool. It had a direct impact on activity recognition as some pairs of activities can finally possess the same image signatures.

One possibility would have been to use surgical tool detection for surgical phase recognition also. The difficulty of this method resides in the fact that most surgical tools are used in every phase, which would have overloaded the image signatures. This information is not of major importance for the recognition of surgical phases, explaining why it was only used for activity detection.

In order to further increase the accuracy of our surgical tool detection method, a solution could be to expand the training image database. Instead of using 100 training images for each surgical tool, we could use dynamic learning through a cross-validation approach. The number of images for each

surgical tool could significantly increase. The best solution for this type of method would be to manually create a huge database with more than 500 images per class. Another way would be to improve the segmentation step by injecting *a priori* knowledge of shapes. The segmentation process is the key of the BVW approach. With perfect segmentation of ROIs containing the surgical tools, the BVW approach should be very accurate. Segmentation methods used in traditional image processing problems, such as graph-cut algorithm, could be further investigated.

VI.3.c. Knowledge-based classification

The strategy consisting in using image signatures composed of visual cues with a constraint on the phases appears promising. Validation studies show modest results (65%) compared to the surgical phase detection results (94%), though these results should be contrasted with the high number of activities (25 possible pairs of activities instead of 8 phases) and the non-sequential aspect of the activities. Indeed, by comparison to phase detection with only 8 classes appearing sequentially only once in the video, many activities appeared more than 5 times during surgery, complicating the recognition task.

The modest value of specificity (54.9%) shows that the system is not very efficient for detecting true positive results, i.e. when activities occur in a frame. On the contrary, it is easier to identify when no activities occurs in the frames (sensitivity = 76.3%). The low value of specificity can be explained by the fact that it is not a binary class problem with just a positive or negative result. Moreover, seven classes were defined for detecting the surgical tools and the assignment of one class for the background allows a good description of this class (using a set of multiple possibilities) in the training database. Using these accurate image signatures, it is therefore easier to identify when no activity occurs during the classification process.

As we can see in **Figure 63**, there are high disparities of recognized and non-recognized frames for each pair of activities. For 5 activity pairs, the numbers of frames over the entire data-set are quite low and the detection is almost null. These 5 pairs are: *<incise, 1.4mm knife, cornea>*, *<aspirate, aspiration cannula, lens>*, *<implant, IOL, anterior chamber>*, *<irrigate, irrigation cannula, conjunctiva and cornea>*, *<wash, irrigation cannula, conjunctiva and cornea>+<irrigate, irrigation cannula, lens>*. Those pairs are due to the way we performed the ground truth. For instance, *<incise, 1.4mm knife, cornea>* is usually paired with *<hold, colibri tweezers, bulbus oculi>*, but this last activity was usually stopped a short amount of time before *<incise, 1.4mm knife, cornea>*. These pairs can therefore be considered as side effects. For other disparities, we can see that some activity pairs are better recognized than others. This is mainly due to the recognition of the surgical(s) tool(s) involved in the dedicated triplets. Moreover, the two pairs of activities that often appear during the surgery (i.e. *<aspirate, syringe, anterior-chamber>+<microspatula, hold, bulbus-oculi>* and *<aspirate, aspiration cannula, lens>+<hold, microspatula, bulbus oculi>*) have recognition rates higher than 60%, probably due to the association of the two surgical tools within the pair that is unique over the set of possible pairs, making easier the recognition.

This combination of human knowledge and data coming from different granularity levels remains the best solution for passing from video images to semantic information. Introducing *a priori* formalized knowledge for the automatic recognition of surgical activities seems to be mandatory, as image-based approaches remain limited to low-level task detection. Moreover, creating only one image signature from a set of high-level features obtained from different algorithms is complex and

may not be optimized. One perspective of this work would be to create a more complex probabilistic model combining all the above information. Moreover, one improvement of the system could be the use of the surgeon's laterality as knowledge before the classification. As surgical tools are used either with the left or right hand, models could be designed to integrate this aspect according to the laterality of the operating surgeon. At the moment, our system does not include this information and we did not assign a side to the activities that we detected.

Additionally, as presented in Chapter II, SPMs using a strong formalization at surgical activity level may help. Such models can be more complex than our hierarchical decomposition making the link between low-level and high-level surgical tasks. Indeed, inside a particular phase, activities can have specific activity sequences where we can extract motifs. For such purpose, motif discovery could therefore be used for finding sequence motifs using computer-based techniques of sequence analysis. In such applications, the motifs are first unknown, and the final results after the analysis are the motifs contained in sequences. During the activity detection step, these results can be used for restraining the possibilities with a more detailed description than we did with our model. This solution would offer the possibility to add, via reasoning on sequence motifs, a strong semantic meaning.

References

- ✓ Agarwal S, Joshi A, Finin T, Yesha Y, Ganous T. A pervasive computing system for the operating room of the future. Mobile Networks and Applications. 2007; 12(2,3): 215-28.
- ✓ Beauchemin SS, Barron JL. The computation of optical flow. ACM New York, USA. 1995.
- ✓ Cramme K and Singer Y. On the Algorithmic Implementation of Multi-class SVMs. JMLR. 2001.
- ✓ Houliston BR, Parry DT, Merry AF. TADAA: Towards automated detection of anaesthetic activity. Methods of Information in Medicine. 2011; 50(5): 464-71.
- ✓ Kragic D, Hager GD. Task modelling and specification for modular sensory based human-machine cooperative systems. Intellignent robots and systems. 2003; 3: 3192-7.
- ✓ Laptev I, Lindeberg T. Local descriptors for spatio-temporal recognition. Spatial coherence for visual motion analysis. Springer. 2006;
- ✓ Miyawaki F, Masamune K, Suzuki S, Yoshimitsu K, Vain J. Scrub nurse and timed-automata-based model for surgery. IEEE Industrial Electronics Trans. 2005; 5(52): 1227-35.
- ✓ Speidel S, Sudra G, Senemaud J, Drentschew M, Müller-stich BP, Gun C, Dillmann R: Situation modeling and situation recognition for a context-aware augmented reality system. Prog Biomed Optics Imaging. 2008; 9(1): 35.
- ✓ Yoshimitsu K, Masamune K, Iseki H, Fukui Y, Hashimoto D, Miyawaki F. Development of scrub nurse robot (SNR) systems for endoscopic and laparoscopic surgery. Micro-NanoMechatronics and Human Science. 2010; 83-8.

Chapter VII. General discussion

In this Chapter, we propose a general discussion about the research performed in this thesis based on the main aspects of the SPM methodology classification that we proposed in Chapter II. From the five aspects that composed a SPM methodology, the analysis methods as well as the validation part have been widely discussed all along the thesis within the specific discussions at the end of each Chapter. We therefore focused our general discussion here on the data acquisition, the modelling and the application aspects. We first discuss about the data acquisition process and propose a discussion on the advantages of using microscope videos as the only sensor from the OR. Then, we address the modelling aspect of this work and especially the level of granularity as well as the formalization that we used all along the recognition process. Finally, we present possible clinical applications that could be used in clinical routine with both the recognition of high- and low- level tasks.

VII.1. Data acquisition

In the continuity of previous works on SPMs, this thesis has been conducted with the motivation of creating automatic recognition tools allowing automatic data acquisition solutions for SPM-based systems. Such data acquisition systems when operational would help the design of a new generation of CAS systems based on SPM. For this objective we decided to use microscope videos data as the only source of information for recognition. This data acquisition solution can be categorized as being low-level information (video), with the surgeon operating intra-operatively, and the method for recording is an on-line video- (**Table 17**).

Table 17 - Classification of our data acquisition technique

	Data acquisition			
	Granularity level	**Operator +/- body part**	**Moment of acquisition**	**Method for recording**
Our work	Low-level (video)	Surgeon	Intra-operative	Video-based recording (on-line)

This choice has been motivated by different criteria. Indeed, as explained by Bouarfa et al. (2010), the information extracted from the OR must be "discriminant, invariant to task distortion, compact in size and easy to monitor". Microscope video data turns out to meet all of these constraints.

Firstly, image features are very discriminant for binary cues extraction, as studies on supervised classification indicated for both datasets, with best accuracies reaching 95% and worst accuracy at 87%. Performance is closely linked to the diversity and the power of discrimination of the databases. That is why the training phase has a major impact on the classification process and must not be ignored.

Secondly, even though it has to be rigourously demonstrated, it seems that within a same surgical environment, procedures are reproducible and image features are thus invariant to task distortion. This

constraint addresses the issue of system adaptability that we addressed in subsection V.7.b. However, due to the different materials and equipment in each department, discriminant images features may differ and the system may be not flexible. For instance, the colour of surgical tissue in Rennes may be different elsewhere and the corresponding features would completely affect the training process. As already mentioned, the solution would be to train dedicated image databases for each surgery that would be adapted to the corresponding surgical environment and microscope scene layout. Using image features for the recognition also eliminates the need to train a dedicated image database for each surgeon. The performance differences between surgeons can be seen in the way that they operate, looking at the hand's motion and dexterity. Assuming that surgical instruments are identical in each hospital, one image database would be sufficient for multiple surgeons. Other variability factors within a dataset can also affect recognition. Ideally, one database should be created for each type of surgery associated with each surgical technique. As a result of this reproducibility, there was no need to use multi-level image annotation for better surgery differentiation, which was done, for instance, by Mueen et al. (2008) for medical image retrieval. In their work, the first step of their multi-level system consisted in capturing the semantic differences between images before more precise annotation. This prior classification is not helpful in our case.

The third crucial parameter is the sample size of acquired data, which must be compact. Image signatures, after data dimension reduction, were composed of 40 features and were thus sharply reduced. Even with modifications in feature selections for another type of surgery, the sample size always stays constant and small. The computation time of the recognition process for one image (feature extraction + data transformation + classification) was 0.8s on a 2-Ghz machine. We did not take into account the computation time of the learning database, considering that this was done off-line.

Lastly, the real value of this project lies in the ease of use of the microscope. Not only is this device already installed in the OR, but also it does not need to be controlled by the surgical staff. It is thus very simple to monitor, enabling the system to be introduced in every department that currently owns a surgical microscope. The use of sensor-based systems that automate the whole recognition process is not restrained to the use of microscope video. The use of global-view videos, sensors positioned on instruments (e.g. RFID tags) or tracking systems has been already tested, validated, and could be a good complementary source of information for the creation of recognition systems. The multiplicity of available sensors in the OR as well as the use of automatic recording of data are the two aspects that need to be further investigated for improving recognition systems.

VII.2. Modelling

As mentioned in Chapter II, the whole SPM methodology is organized around the aspect of granularity level. If we position this work compared to the proposed granularity axis of **Figure 6**, the modelling here focused on a granularity level between the activities and the phases. Adding for instance a source from a global-view video would allow handling the granularity level following the surgical phases, i.e. the surgical procedure. On the contrary, adding a source of information like tracking systems on instrument could allow the modelling handling a lower granularity level, i.e. motions.

It's also necessary for *bottom-up* approaches, approach that we follow in this thesis, to have a formal representation of surgery. The formalization that we used in Chapter I for recognizing high-level tasks was very light and was represented as a simple state-transition diagram (**Table 18**).

Table 18 - Classification of our modelling

	Modelling		
	Granularity level	**Operator +/- body part**	**Formalization**
High-level tasks	Phases	Surgeon	State-transition diagram
Low-level tasks	Activities	Surgeon	Hierarchical decomposition

Positioning this formalization on the axis of **Figure 7** shows that it remains a light formalization that could be further developed. For the detection of low-level tasks, however, we used a more complex formalization, i.e. a hierarchical decomposition (**Table 18**). This was required to go down from one level and detect more details of the surgical procedure. The message here is that formalization is needed for being able to correctly recognized high- and low-level tasks of a surgical procedure. Moreover, it can be also necessary for comparing and sharing studies between different centres. A heavy and rich formalization is therefore the key of future analysis of SPMs, for helping the recognition process as well as for easy sharing between centres.

VII.3. Clinical applications of the developed frameworks

The two main applications of the frameworks we developed are following (**Table 19**). Surgical videos are increasingly used for learning and teaching purposes, but surgeons often do not use them because of the huge amount of data in surgical videos (hours of video to be browsed) and the lack of data organisation and storage. It could therefore be introduced into clinical routine for post-operative video indexation and creation of pre-filled reports. One can imagine a labelled database of videos with full and rapid access to all surgical tasks for easy browsing. The video database created would contain the relevant surgical tasks of each procedure for easy browsing. We could also imagine the creation of post-operative reports, automatically pre-filled with recognised events that will have to be further completed by the surgeons themselves. For such clinical applications, even with few errors and an off-line use, automatic indexation would be relevant, as there is no need for perfect detection and it has no impact on the surgery itself. In its present form, the computation time of the recognition process for one frame (visual cues detection + DTW/HMM classification) was evaluated to around 3s on a standard 2-Ghz computer, which makes possible the development of off-line applications. However, we did not implement any of these applications. We are therefore not able to demonstrate the feasibility or the added value brought by our methods. With the present methodology, however, both frameworks for the recognition of high-level and low-level tasks could only be introduced into the surgical routine as assistance for off-line applications only. The DTW algorithm requires knowing the entire performed procedure to determine the optimum path, avoiding on-line use. For on-line applications, the HMM classification should therefore be used.

Once on-line automatic recognition approaches will be developed, the explicit computer based understanding of OR high-level and low-level tasks could help developing an intelligent architecture that analyses microscope images and transforms them into a tool for assisting the decision-making process. It could also support intra-operative decision making by comparing situations with previously recorded or known situations. This would result in a better sequencing of activities and improved anticipation of possible adverse events, which would, on the one hand optimize the surgery, and on the

other hand improve patient safety. Such context-awareness feature would also be very useful for improving ergonomy of the systems in the OR, for instance, by specifying which kind of information needs to be displayed for the surgeon's current task. This would also help to improve OR management (as in Xiao et al., 2005) and support CAS systems (as in Jannin et al., 2007; Speidel et al., 2008).

Table 19 - Classification of our clinical applications

	Application	
	Surgical speciality	**Clinical application**
High-level surgical tasks	Neurosurgery – Hypophyse Eye surgery - Cataract	Intra-operative assistance Training/assessment of surgeons
Low-level surgical tasks	Eye surgery - Cataract	Intra-operative assistance Training/assessment of surgeons

References

- ✓ Bouarfa L, Jonker PP, Dankelman J. Discovery of high-level tasks in the operating room. J Biomed Inform. 2010; 44(3): 455-62.
- ✓ Jannin P, Morandi X. Surgical models for computer-assisted neurosurgery. NeuroImage. 2007; 37(3): 783-91.
- ✓ Speidel S, Sudra G, Senemaud J, Drentschew M, Müller-stich BP, Gun C, Dillmann R. Situation modelling and situation recognition for a context-aware augmented reality system. Progress Biomed Optics Imaging. 2008; 9(1): 35.
- ✓ Xiao Y Hu P, Hu H, Ho D, Dexter F, Mackenzie CF, Seagull FJ. An algorithm for processing vital sign monitoring data to remotely identify operating room occupancy in real-time. Anesth Analg. 2005; 101(3): 823-32.

Chapter VIII. Conclusion

Devant l'apparition de nombreux outils et nouvelles technologies dans les salles d'opération, le besoin de nouveaux systèmes de CAO se fait actuellement ressentir. La modélisation du processus chirurgical semble être un des élèments clés pour la construction de la salle d'opération du futur. En d'autres termes, un des challenges de le CAO est d'assister la procédure à travers la compréhension des évènements de la salle d'opération. C'est pourquoi une des branches des SPMs la plus étudiée jusqu'à présent est la reconnaissance de tâches haut et bas-niveaux dans les salles d'opération via des méthodes appelées méthodes *bottom-up*. Ces approches utilisent des signaux enregistrés par des capteurs ou par des humains dans la salle d'opération pour reconnaitre automatiquement des informations de plus haut niveau, telles que les phases, étapes, activités de la procédure chirurgicale. Devant ces avancées récentes, nous avons proposé dans cette thèse d'utiliser les vidéos des microscopes chirurgicaux comme unique source d'information afin de créer des systèmes automatiques de reconnaissance de tâches chirurgicales.

Les raisons d'utiliser des signaux vidéo, et spécifiquement des signaux vidéo issus des microscopes, en entrée sont doubles. Premièrement, cela permet de collecter des informations sur la procédure sans pour autant altérer la routine clinique ni gêner le chirurgien. Devant le nombre important de capteurs déjà installés, il n'est pas nécessaire de surcharger cette salle. Cela amène à réfléchir à des solutions simples de contrôle et surveillance que peuvent être les vidéos. De même, comme vu dans le Chapitre II, de nombreuses études du domaine se sont déjà basées sur des capteurs ajoutés dans les salles d'opérations et qui pouvaient, dans des cas bien particulier, altérer le processus chirurgical. Deuxièmement, l'enregistrement de la vidéo est courante, standardisé, et permet de rendre le processus de reconnaissance entièrement automatique. A partir de ces vidéos de microscopes chirurgicaux, nous avons proposé deux types d'approches permettant la reconnaissance automatique des tâches chirurgicales se focalisant toutes les deux sur des niveaux de granularité distincts. Le premier système a été créé pour reconnaitre des tâches haut-niveaux, symbolisées par les différentes phases d'une procédure chirurgicale. Le deuxième système est descendu d'un niveau de granularité pour reconnaitre des tâches de bas-niveau, symbolisées par les activités détaillées du chirurgien.

La première partie de la thèse s'est donc focalisée sur la reconnaissance automatique des phases chirurgicales. L'idée sous-jacente fut de combiner des techniques de vision par ordinateur robustes permettant d'extraire des attributs visuels, avec une analyse de séries temporelles pour prendre en compte l'aspect séquentiel des phases. Premièrement, des attributs visuels pertinents qui permettent de discriminer les différentes phases de la chirurgie furent manuellement définis. Cinq classifieurs furent implémentés pour reconnaitre ces attributs dans l'image, chacun étant relié à un type de caractéristiques à extraire. Les attributs reconnaissables à travers leurs couleurs furent extraits avec des histogrammes de couleurs. Pour les attributs reconnaissables à travers leurs formes, deux types de classifieurs furent implémentés. Le premier fut un classifieur de Haar pour catégoriser des objets à forts contours. Le deuxième fut une approche par sac-de-mots pour détecter des objets sans les catégoriser. Les attributs reconnaissables à travers leurs textures furent appréhendés par une approche

par sac-de-mot, et enfin tous les autres attributs ne rentrant pas clairement dans un de ces types de caractéristiques furent reconnus grâce à un classifieur standard mêlant extraction de caractéristiques spatiales, sélection de caractéristiques et classification supervisée. Cette première étape de traitement d'image pur permet de caractériser chaque frame de la vidéo de façon statique, et après concaténation des signatures images de créer des séries temporelles qui peuvent être ensuite présentées en entrée de système d'analyse de séries temporelles. Nous avons implémenté deux de ces méthodes, les *chaînes de Markov Cachées* et l'algorithme *Dynamic time Warping*.

La seconde partie de la thèse s'est focalisée sur la reconnaissance automatique des activités chirurgicales (tâches bas-niveau). Ce niveau de granularité est formalisé par des triplets < *action – outil chirurgical – structure anatomique* >. Des informations plus précises sur les outils chirurgicaux ainsi que sur les zones d'apparitions de ces outils furent mixées avec les attributs visuels précédemment extraits au sein de signatures image plus détaillées. Ensuite, en se basant sur l'hypothèse que la plupart des activités apparaissent seulement dans une ou deux phases, une décomposition hiérarchique de la procédure fut créée pour faire le lien entre phase et activité. En utilisant cette décomposition hiérarchique, les résultats de la classification des phases ainsi que les nouvelles signatures images, la reconnaissance des activités devient possible.

Les études de validation ont été menées sur deux jeux de données très différents : un jeu de 16 vidéos de chirurgie hypophysaire, qui est un cas particulier de tumeur neurochirurgicale, et un jeu de 20 vidéos de chirurgie de la cataracte, qui est une chirurgie courante de l'œil. Dans le cas de la chirurgie de l'hypophyse, des taux de reconnaissance de phases de l'ordre de 92% ont été obtenus. Dans le cas de la chirurgie de la cataracte, des taux de reconnaissance de 94% ont été obtenus pour la détection de phases, et de l'ordre de 65% pour les activités. Ces taux de reconnaissance permettent d'effectuer la modélisation des chirurgies en identifiant l'enchainement des différentes phases et activités.

En ce qui concerne la reconnaissance des phases, le système est très performant et les résultats obtenus sont très convaincants. Pour la reconnaissance des activités, notre système offre des premiers résultats prometteurs. En effet, des améliorations peuvent être apportées. Notre architecture avec plusieurs niveaux de granularité (phases, activités) permet d'être modulable et de fournir des informations encore plus précises en fonction de la situation. Par exemple, en fonction des installations présentes dans la salle d'opération, il peut être possible d'enrichir les différentes signatures images avec des informations provenant d'autres senseurs, de type image ou non image. De même, nos méthodes de vision par ordinateur et d'analyse de séries temporelles, utilisées dans cette thèse sur des vidéos de microscope, sont tout-à-fait adaptables à d'autres types de vidéos, comme des vidéos grands champs de salle d'opération ou des vidéos focalisées sur un acteur spécifique de la chirurgie.

Ces systèmes de reconnaissance de tâches chirurgicales, que ce soit au niveau des phases ou au niveau des activités, apparaissent comme une progression non négligeable vers la construction de systèmes intelligents (autrement dit sensibles au contexte) pour la chirurgie. Dans leurs versions actuelles, les systèmes peuvent être utilisés de manière postopératoire afin d'indexer les vidéos en fin de chirurgie et de créer des rapports chirurgicaux pré-remplis. Dans le cadre de l'enseignement, avoir à disposition une base de données de vidéos chirurgicales indexées peut être aussi utile et une navigation entre les différentes phases et activités des chirurgies pourrait être effectuée. Une des perspectives principales de cette thèse est l'utilisation de systèmes équivalents dans les salles d'opération en temps-

réel. Pour le moment, certain algorithmes (DTW par exemple) ne fonctionnent uniquement lorsque la vidéo est entièrement terminée, ce qui limite les champs d'application du système. Vers cet objectif, une des applications temps-réel qui pourrait être amenée à voir le jour est l'assistance intra-opératoire, par exemple en permettant en temps réel de savoir quelles informations ont besoin d'être montrées au chirurgien pour la tache effectuée. Cela pourrait aussi permettre une meilleure anticipation de possibles évènements néfastes permettant d'une part d'optimiser la chirurgie et d'autre part de réduire les dangers pour le patient. Les systèmes de reconnaissances basées sur les vidéos des microscopes, que ce soit pour la détection des tâches de haut ou bas-niveau, offrent donc de réelles perspectives d'avenir dans le domaine de la CAO.

Appendix A – ICCAS editor software

The ICCAS editor software allows recording surgical procedures directly in the OR or after the intervention on videos at the activities granularity level. It records activities with the same formalization that we presented in subsection III.2.c. A structured .xml file is generated at the end of the recording containing all needed information (start time and end time of each activity, meta-information, etc...).

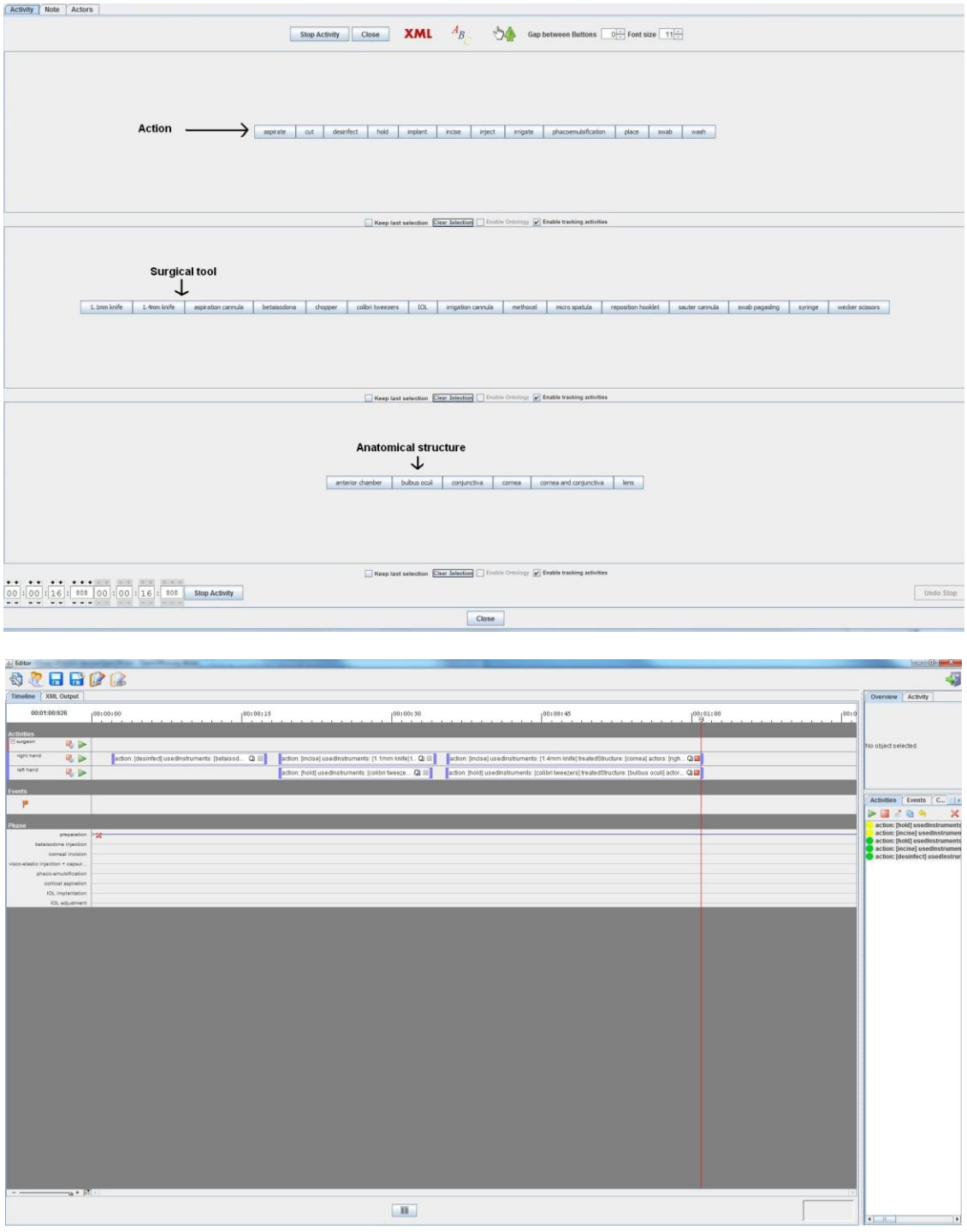


Figure 64 - Screenshots of the ICCAS editor software

Appendix B – C++/Qt GUI

For the needs of the project, a graphical user interface (GUI) in C++/Qt was implemented by David Bouget (see Software development) in order to perform live experiments of image processing techniques. This software allows the user to modify a large set of parameters over the entire recognition framework. In particular, it includes a display and test mode, a spatial interface, a temporal interface and an interface for the detection of activities (**Figure 65**). This software was very helpful for optimizing all parameters and variables all along the project.

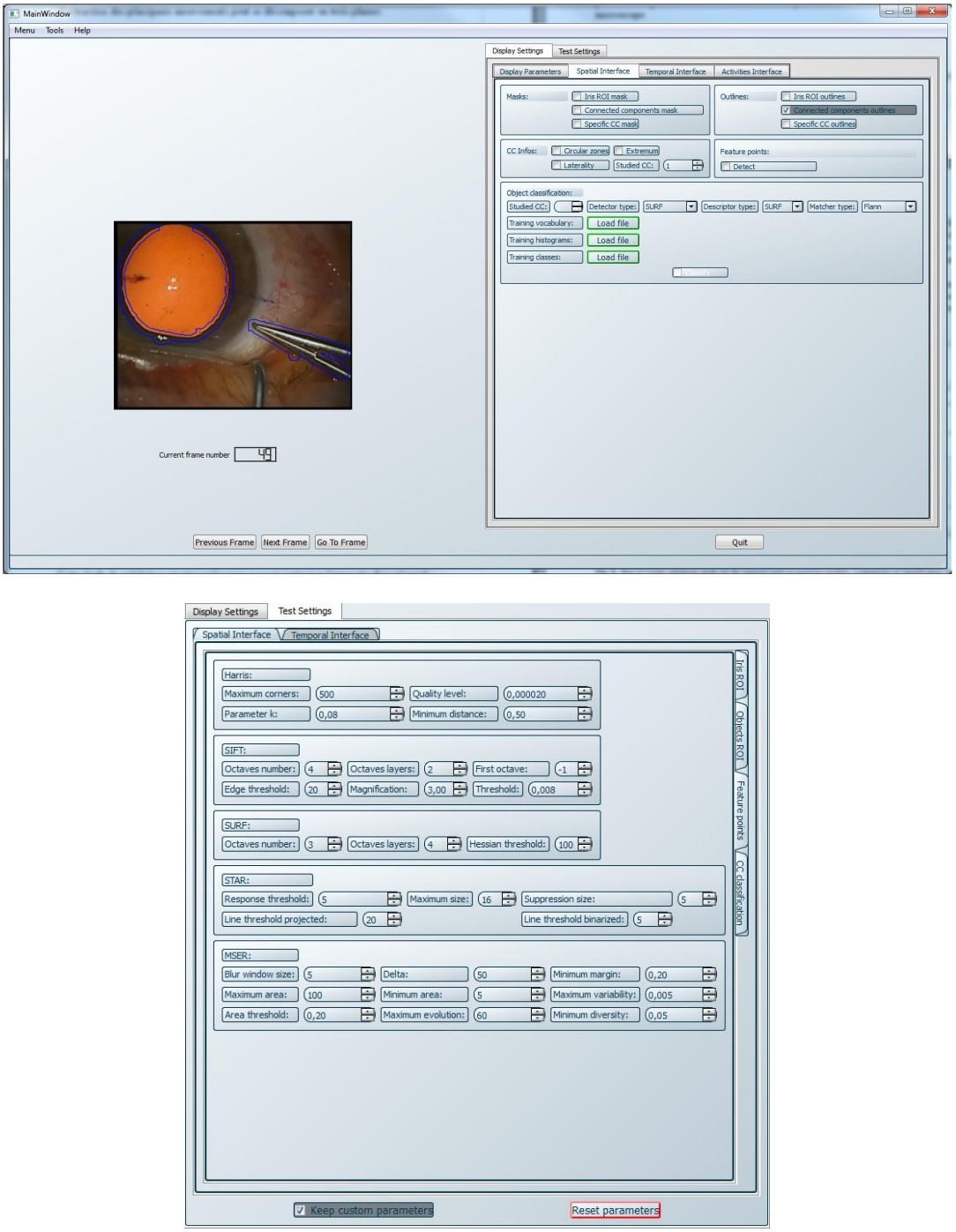


Figure 65 - Screenshots of the GUI. Display mode (above) and test mode (below) of the spatial interface.

Appendix C – Matlab GUI

Also for the needs of the project, a GUI in Matlab was implemented. The system is based on two development platforms, C++ and Matlab. The C++ part is used for imaging processing, and the Matlab part for the time-series algorithms. The feature extraction step has been done off-line, and then integrated to the GUI. The Demo is indeed an off-line application that has been created only to show the perspective of the system and to test potential combination of different time-series parameters. When launching the demo application, the user could choose among different parameters:

- ✓ The type of image features that will compose the image signatures. The RGB and HSV spaces (color), the co-occurrence matrix (texture), the Hu moments and the DCT transform (form) are the features available. Three different combinations of these features can be chosen through the GUI: Color / Color + Texture / Color + Texture + Form.

- ✓ The number of videos for the learning stage. A total of 20 videos of cataract is available. If x videos is chosen by the user for the learning, then the 20-x others will be used for the validation.

- ✓ The type of time series approach for the modeling, choosing between HMM and DTW algorithms. When the user chooses the DTW algorithm, he can also choose a global constraint on the alignment sequence (Sakoe-Chiba constraint or Itakura parallelogram constraint). The parameters of both global constraints can be modified.

Figure 66 shows a screen-shot of the demo. The user first has to choose the type of image features that he wants for the detection, then he has to load the database using the desired number of videos for training, and lastly, using a chosen method for modeling. All validation videos are processed and displayed.

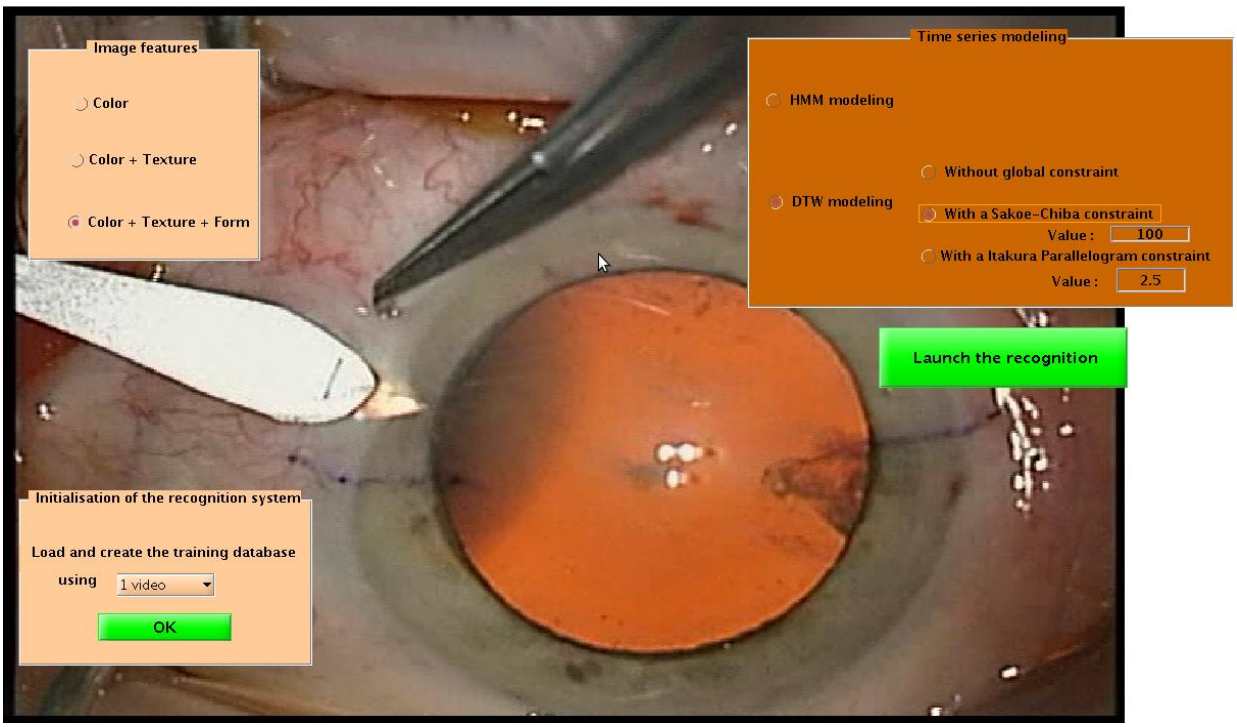


Figure 66 - Screen-shot of the Matlab GUI

Appendix D – Publications

Journals

2012 – A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. **Lalys F**, Riffaud L, Bouget D, Jannin P. Trans on Biomedical Engineering. 59(4), p. 966-976.

2012 – Automatic knowledge-based recognition of low-level tasks in the OR. **Lalys F**, Bouget D, Riffaud L, Jannin P. Int J Comput Assist Radiol Surg (*ahead of print*)

2012 – Classification of Surgical Processes using Dynamic Time Warping. Forestier G, **Lalys F,** Riffaud L, Trelhu B, Jannin P. J Biomed Inform. 45, p. 255-264.

2011 - Automatic computation of electrode trajectories for Deep Brain Stimulation: a hybrid symbolic and numerical approach. Essert C, Haegelen C, **Lalys F**, Abadie A, Jannin P. Int J Comput Assist Radiol Surg (*ahead of print*)

2011 - Construction and assessment of a 3-T MRI brain template. **Lalys F**, Haegelen C, Ferre JC, El-Ganaoui O, Jannin P. NeuroImage. 49 (1), p. 345-354.

Conferences with proceedings

2012 – Surgical tools recognition and pupil segmentation for cataract surgery modeling. Bouget D, **Lalys F**, Jannin P. *MMVR, Newport Beach, United States*. Stud Health Technol Inform. IOS press book. 173, p. 78-84.

2012 – Analysis of electrodes' placement and deformation in deep brain stimulation from medical images. Mehri M, **Lalys F**, Maumet C, Haegelen C, Jannin P. *SPIE medical imaging*, 8316-32. San Diego, United States.

2012 – Clustering de sequences d'activités pour l'étude de procedures neurochirurgicales. Forestier G, **Lalys F,** Riffaud L, Trelhu B, Jannin P. *EGC*. p. 489-494 Bordeaux, France.

2011- An application-dependent framework for the recognition of high-level surgical tasks in the OR. **Lalys F**, Riffaud L, Bouget D, Jannin P. *MICCAI*. 14(1), p. 331-339. Toronto, Canada.

2011- Analyse de vidéos de microscopes chirurgicaux pour la reconnaissance automatique d'étapes en combinant SVM et HMM. **Lalys F**, Riffaud L, Morandi X, Jannin P. *ORASIS*, Praz-Sur-Arly.

2011- Correlating Clinical Scores with Anatomical Electrodes Locations for Assessing Deep Brain Stimulation. **Lalys F**, Haegelen C, Abadie A, Jannin P. *IPCAI*, Berlin, Germany. 6689, p. 113-121.

2010 - Surgical Phases Detection from Microscope Videos by Combining SVM and HMM. **Lalys F**, Riffaud L, Morandi X, Jannin P. *MCV 2010 (MICCAI Workshop)*, Beijing, China. p. 54-62.

2010 - Automatic phases recognition in pituitary surgeries by microscope images classification. **Lalys F**, Riffaud L, Morandi X, Jannin P. *IPCAI*. Geneve, Switzerland. p. 34-44.

2009 - Post-operative assessment in Deep Brain Stimulation based on multimodal images: registration workflow and validation. **Lalys F**, Haegelen C, Abadie A, Jannin, P. *SPIE Medical Imaging*. Lake Buena Vista, FL, United States.

2009 - Analyse de données pour la construction de modèles de procédures neurochirurgicales. Trelhu B, **Lalys F**, Riffaud L, Morandi X, Jannin P. *EGC*. Strasbourg, France. p. 427-432

2009 - Analyse post-opératoire en stimulation cérébrale profonde basée sur des images multimodales : mise en place et validation des chaînes de recalage. **Lalys F**, Haegelen C, Abadie A, El Ganaoui O, Jannin P. *ORASIS*. Trégastel, France.

Conferences/workshops without proceedings

2011 - Assessment of surgical skills using Surgical Processes and Dynamic Time Warping. Forestier G, **Lalys F**, Riffaud L, Trelhu B, Jannin P. *M2CAI 2011 (MICCAI workshop)* - Toronto, Canada

2011 - Validation of basal ganglia segmentation on a 3T MRI template. Haegelen C, Guizard N, Coupé P, **Lalys F**, Jannin P, Morandi X, Collins DL. *Human Brain Mapping*, Quebec City, Canada

2010 - Validation de la Segmentation des Ganglions de la Base sur un Template IRM 3 Tesla. Haegelen C, **Lalys F**, Abadie A, Collins L, Brassier G, Morandi X. *Réunion de la Société de Neurochirurgie de Langue Française*, Marne-la-Vallée, France.

2010 - Anatomo-clinical atlases in subthalamic Deep Brain Stimulation correlating clinical data and electrode contacts coordinates. **Lalys F,** Haegelen C, Baillieul M, Abadie A, Jannin P. *IBMISPS*, United States.

Software

2012 - ProcSide: software for recording surgical procedures. *Dépôt APP*. Bouget D, Jannin P, **Lalys F**.

Résumé étendu de la thèse

I. Introduction

I.1 Contexte

Une forte augmentation de nouvelles technologies dans les systèmes de santé se fait actuellement ressentir, allant du management et de l'organisation de l'hôpital jusqu'aux solutions d'imagerie médicale. Dans les salles d'opération, les technologies de l'informatique sont maintenant essentielles et de plus en plus utilisées tout au long de l'intervention chirurgicale : du planning pré-opératoire à l'évaluation post-opératoire, en passant bien sûr par l'aide intra-opératoire. C'est dans ce contexte que sont nés les systèmes de Chirurgie Assistée par Ordinateur (CAO). La CAO est définie comme l'ensemble des systèmes aidant le praticien dans la réalisation de ses gestes diagnostiques et thérapeutiques, ou plus simplement par des procédures chirurgicales conduites à l'aide d'ordinateurs. Il est en effet important de concevoir une salle d'opération qui offre au chirurgien et à son équipe une facilité de travail et d'accès aux images, informations et outils disponibles. Notamment, la modélisation du processus chirurgical, c'est-à-dire du déroulé de l'intervention, est une information importante vers la construction de la salle d'opération du futur. Une procédure chirurgicale est décrite de manière principalement symbolique comme une succession d'étapes et d'actions réalisées avec différents outils et selon différentes techniques. Les modèles pouvant prendre en compte ces différents paramètres semblent ainsi être la base des nouveaux systèmes de CAO autour desquels s'inscrit cette thèse. Nous allons donc commencer par introduire les modèles de processus chirurgicaux, i.e. *Surgical Process Model* (SPM), puis nous introduirons la problématique de cette thèse.

I.2 Etat de l'art en SPM

Un même type de procédure chirurgicale est reproductible, et cette hypothèse rend possible une modélisation temporelle des procédures, dont le but est de collecter des données et de créer des modèles issus de ces données (Jannin and Morandi, 2007). Pour étudier les nombreuses publications couvrant le domaine, nous avons proposé une classification basée sur 5 aspects de la méthodologie de création des SPMs : l'acquisition de données, la modélisation, l'analyse, les applications et la validation. Une revue de la littérature basée sur une recherche Google Scholar fut effectuée, et 43 papiers furent ainsi sélectionnés, rapportés et classés. Parmi ces 5 aspects, l'aspect d'acquisition de données sert de base aux modélisations. Les méthodes d'acquisition qui en découlent suivent deux stratégies différentes : acquisition par des opérateurs humains positionnés dans la salle d'opération, ou par des capteurs de façon automatique. Les approches basées « opérateur humain » ont la capacité de couvrir des niveaux de granularité supérieurs aux approches basées « capteurs » en incluant des informations sémantiques. Cependant, ces approches « opérateur humain » présentent la limite de ne pas être automatisées, et donc de nécessiter beaucoup de temps et de ressources humaines. C'est pourquoi les approches basées « capteurs » sont de plus en plus étudiées pour automatiser ce processus d'acquisition de données. Au sein de ces approches, on retrouve les études utilisant des simulateurs ou environnements virtuels complets (Darzi et al., 2002 ; Lin et al., 2006), pour étudier la gestuelle des chirurgiens et créer des modèles de reconnaissance de gestes. On retrouve également les systèmes de reconnaissance basés sur des capteurs installés sur les instruments (Ahmadi et al., 2007 ; Padoy et al.,

2007). Des informations binaires de présence d'instruments sont analysées et des modèles graphiques probabilistes sont utilisés pour reconnaître des informations de haut niveau dans la salle d'opération. D'autres systèmes ont aussi été testés, comme la mise en place d'un outil de suivi du regard du chirurgien (James et al., 2007), ou d'un système de tracking 3D de la positon de chaque membre du staff (Nara et al., 2009). Le principal défaut de tous ces systèmes est qu'ils ne sont pas installés d'office dans les salles d'opération, et que la mise en place de trop nombreux outils pourrait à long terme gêner le déroulé de l'intervention.

Pour pallier à ce problème, les nouvelles études se focalisent sur des sources d'informations déjà installées dans la salle d'opération, telles que les signes vitaux des patients (Xiao et al., 2005) ou les vidéos (Bhatia et al., 2007 ; Speidel et al., 2008; Klank et al., 2008; Lo et al., 2003). L'utilisation de vidéos, grand champs ou endoscopiques, permet d'automatiser l'acquisition de données sans altérer la routine clinique. Des outils de vision par ordinateur et de réalité augmentée sont utilisés pour extraire des informations pertinentes au chirurgien. Les images vidéos des différentes caméras se révèlent donc être une source riche en informations pouvant éventuellement remplacer les approches se basant sur des enregistrements humains.

I.3 Problématique de la thèse

Devant l'intérêt croissant des études sur les SPMs, et à partir de la revue de la littérature que nous avons effectuée, cette thèse s'est centrée sur quatre aspects principaux. Premièrement, l'idée fut de développer des nouveaux outils pour la reconnaissance de tâches chirurgicales haut et bas-niveaux. Deuxièmement, nous avons utilisé comme unique source d'information les vidéos des microscopes, utilisés de façon systématique tout au long d'une intervention neurochirurgicale ou ophtalmologique. Troisièmement, différents niveaux de granularités des tâches chirurgicales (i.e. phases et activités) ont été couverts. Enfin, nous avons introduit une sémantique forte à notre modélisation.

II. Jeux de données

Les différentes méthodes que nous avons mises en place tout au long de cette thèse furent testées sur 2 jeux de données différents.

1) La chirurgie des adénomes hypophysaires, type particulier de neurochirurgie. Une voie d'abord transnasale est utilisée par les chirurgiens, qui atteignent ensuite l'hypophyse pour enlever la tumeur située dans cette région. Nous disposions de 16 vidéos (temps moyen de chirurgie : 50min), où le chirurgien a défini 6 phases (**Figure 1**, gauche).
2) La chirurgie de la cataracte, type de chirurgie ophtalmologique. Le principe est d'enlever la lentille naturelle de l'œil (le cristallin) pour la remplacer par une lentille artificielle. Nous disposions de 20 vidéos (temps moyen de chirurgie: 15min) et huit phases furent identifiées (**Figure 1**, droite). Pour définir les activités, nous nous sommes basés sur la formalisation proposée par Neumuth et al. (2007), décrivant une activité comme un triplet : < *verbe d'action – outil chirurgical – structure anatomique* >. 12 verbes d'action, 13 outils chirurgicaux et 6 zones d'action furent identifiés. Toutes les combinaisons ne sont bien évidemment pas possibles, car n'ayant aucun sens, ce qui a amené à identifier 17 activités puis 25 paires d'activités possibles (une activité par main du chirurgien). Dans ces 25 activités, l'activité « arrière-plan » est une activité à part entière. Un exemple de visualisation par « index-plot » est proposé sur la **Figure 2**.

Pour ces deux jeux de données, des sous-échantillonnages spatiaux et temporels furent effectués comme étape de pré-traitement.

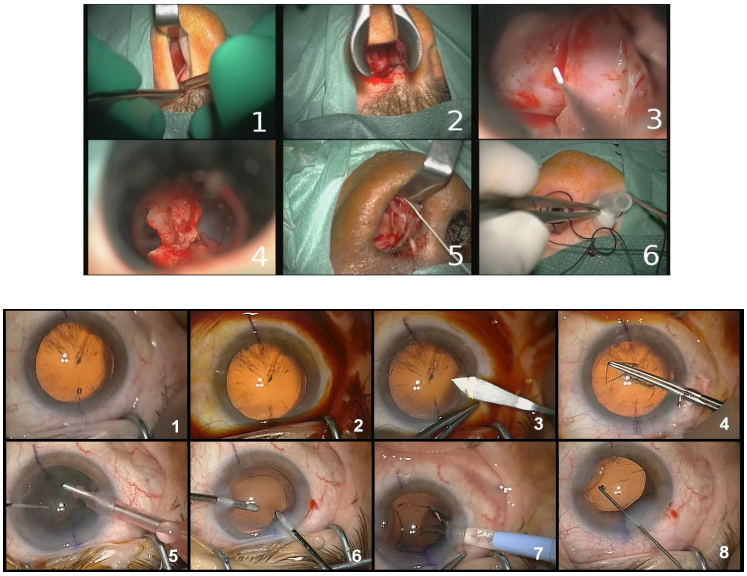


Figure 1 - Exemple d'images des microscopes pour les deux jeux de données :
Haut : Chirurgie des adénomes hypophysaire : 1-incision nasale, 2-installation des écarteurs nasaux, 3-exérèse de la tumeur, 4-remise en place de la cloison nasale, 5-suture, 6-installation des compresses nasales
Bas : Chirurgie de la cataracte : 1-préparation, 2-injection de Bétadine, 3-incision de la cornée, 4-hydrodissection, 5-phakoemulsification, 6-aspiration corticale, 7-implantation de la lentille artificielle, 8-ajustement de la lentille

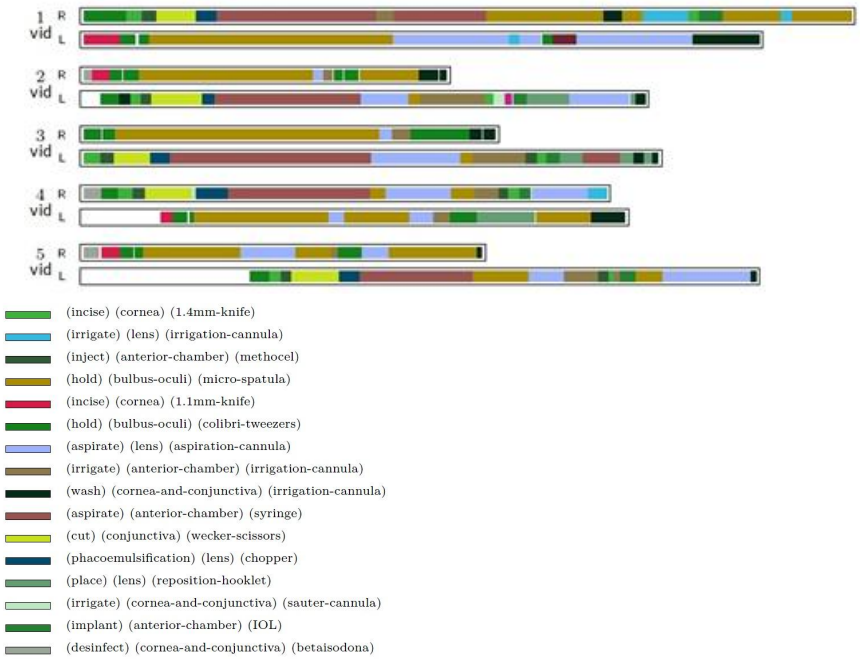


Figure 2. Représentation par « index-plot » des activités pour 5 vidéos de chirurgie de la cataracte

III. Reconnaissance des phases : approche statique

Dans la première partie de cette thèse, nous nous sommes focalisés sur la reconnaissance des phases de façon statique, c'est-à-dire sans prendre en compte l'aspect séquentiel des phases. Chaque image est donc classée indépendamment des autres en suivant une approche traditionnelle de classification d'image.

III.1 Méthodes

Le système de reconnaissance que nous avons suivi pour cette approche est le suivant : après une étape de pré-traitement, une extraction de caractéristiques fut effectuée pour chaque image, suivie d'une sélection des meilleures caractéristiques. Une classification supervisée a ensuite permis d'assigner une phase à chaque image des vidéos. Le système est présenté sur la **Figure 3**.

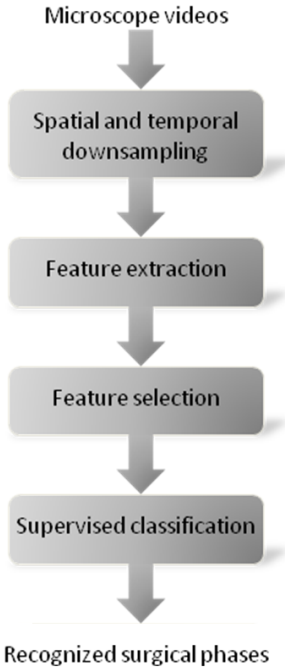


Figure 3 - Système de reconnaissance de phases

Extraction et sélection des caractéristiques

Dans cette étape, nous nous basons sur une analyse de chaque image de vidéo chirurgicale de manière ponctuelle dans le but d'extraire de nouvelles caractéristiques images purement spatiales. Pour chaque image, une signature a été extraite, composée de caractéristiques de texture, de forme et de couleur, dans le but de chercher des similarités entre images de même classe. La couleur a été extraite avec deux espaces complémentaires (Smeulders et al. 2000), l'espace RVB (Rouge Vert Bleu) et l'espace TSV (Teinte Saturation Valeur). Pour la texture, nous avons opté pour les matrices de co-occurrences associées aux descripteurs d'Haralick (Haralick et al., 1973). La forme a été obtenue à partir des moments spatiaux (Hu, 1962). Enfin les coefficients de la Transformée en Cosinus Discrète (TCD) ont été calculés (Ahmed et al., 1974). 185 caractéristiques spatiales furent ainsi extraites. Nous avons ensuite utilisé une méthode hybride de sélection combinant une approche dit *filter* d'une approche dit

wrapper (Duda and Hart, 1973) pour réduire la taille de ces signatures tout en gardant le maximum d'informations pertinentes. La première approche exécute la sélection en regardant les propriétés intrinsèques des données. La deuxième utilise un classifieur en sortie pour évaluer la meilleure combinaison de variables d'entrée. Ces deux types d'approches sont complémentaires et peuvent être fusionnées par l'intersection des résultats des deux algorithmes. L'information mutuelle (Hamming, 1980) et l'algorithme SVM Recursive Feature Elimination (RFE) furent choisis pour chaque approche. En complément de cette approche hybride de sélection, une analyse en composante principale (PCA) fut aussi comparée.

Classification supervisée

Plusieurs techniques classiques de classifications supervisées ont été testées : Machine à Vecteur de Support (SVM), Plus Proche Voisin (PPV), Réseaux neurones (RN), arbre de décision et Analyse Discriminante Linéaire (ADL). Ces algorithmes furent évalués grâce à une validation croisée 10-fold, sur les deux jeux de données séparément. En complément de la probabilité de bonne classification, la sensibilité et la spécificité furent aussi calculées.

III.2 Résultats

La **Figure 4** montre que l'ACP est mieux adaptée pour un nombre de caractéristiques inférieur à 30. A partir de ce seuil, la méthode hybride de sélection donne de meilleurs résultats et atteint son maximum pour 40 caractéristiques. Pour l'ACP, la précision diminue à partir de 40 caractéristiques, alors qu'en utilisant l'autre approche de réduction de dimension, la précision reste pratiquement inchangée.

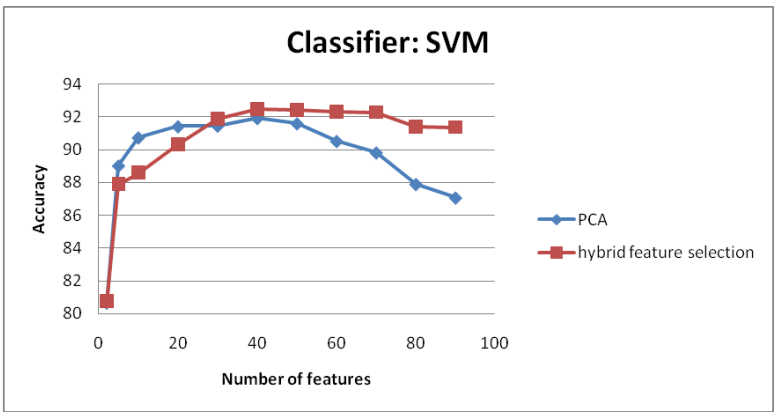


Figure 4 - Probabilité de bonne classification selon le nombre de composantes gardées.

Tableau 1 - Probabilité de bonne classification (Pbc), sensibilité et spécificité pour les 5 algorithmes étudiés et en ne gardant que 40 caractéristiques.

Algorithmes	Pbc	Sensibilité	Spécificité
SVM	82.2%	78.7%	98.1%
PPV	74.7%	66.0%	95.4%
RN	71.3%	65.1%	92.8%
Arbres decision	66.2%	52.3%	94.0%
ADL	81.5%	77.0%	97.6%

Avec 40 caractéristiques, les autres classifieurs sont testés sur le **Tableau 1**. Les SVMs donnent les meilleurs résultats (91.5%), suivis par l'ADl et les PPV. En revanche, les arbres de décision et les réseaux de neurones donnent les moins bons résultats.

III.3 Discussion

Réduction de dimensionnalité : Deux méthodes furent testées pour réduire la dimension des données images : l'ACP et une méthode hybride de sélection de caractéristiques combinant une approche *wrapper* avec une approche *filter*. Intuitivement, les méthodes *wrapper* semblent plus avantageuses, puisqu'elles utilisent les résultats des classifications pour faire la sélection. En revanche, la principale limitation reste le temps de calcul, qui augmente de façon exponentielle avec la taille des données. Pour les approches *filter*, la sélection est faite sans regarder les résultats des classifications, mais seulement sur les données d'entrée, en évaluant le pouvoir prédictif de chaque variable. La principale limite se trouve dans l'incapacité de telles méthodes à prendre en compte des combinaisons de caractéristiques, ce qui affecte la précision de la sélection. L'avantage de la combinaison est donc de bénéficier des avantages des deux méthodes. Une étude de comparaison (**Figure 3**.) de cette approche avec l'ACP a clairement montré la supériorité de cette méthode.

Classification supervisée : La précision des SVMs, associée à une faible déviation standard, a montré la robustesse de cet algorithme pour ce type d'images. Les bonnes performances, en grande partie expliquées par leurs capacités de généralisation, ne sont pas surprenantes si on regarde la récente explosion de son utilisation. L'ADL, même si ses performances se dégradent rapidement, restent aussi une bonne méthode de classification. En revanche, les arbres de décision et l'algorithme des PPV ont montré leurs limites pour nos jeux de données. Ces derniers étaient vraisemblablement trop variables en couleur et en texture et pas assez discriminants pour utiliser des outils de classification simples. Le résultat des réseaux de neurones peut paraitre surprenant, compte tenu du type de données disponibles. Les algorithmes non-linéaires sont généralement adaptés aux systèmes très complexes, en revanche les algorithmes linéaires sont plus faciles et rapides à utiliser, ce qui les rend adaptés à notre problématique.

De l'approche statique à l'approche dynamique : Après des études comparatives menées sur 2 jeux de données issus de vidéos de microscopes neurochirurgicaux, il a été démontré que la reconnaissance automatique des étapes d'une chirurgie était possible avec un taux de bonne classification de l'ordre de 82%. Cependant, des confusions entre étapes distantes ont été repérées, et cette première approche, bien que nécessaire pour appréhender l'étude, n'est pas suffisante pour être intégrée dans des applications cliniques.
La solution envisagée fut l'ajout de l'information temporelle, qui permettra de résoudre une bonne partie de ces confusions et de générer des systèmes plus fiables et robustes. Dans la deuxième partie de cette thèse, nous avons ainsi pris en compte l'aspect séquentiel des phases en utilisant des algorithmes de classification de séries temporelles.

IV. Reconnaissance des phases : approches dynamique

Dans cette deuxième partie de la thèse, nous avons tenté d'améliorer le premier système de reconnaissance de phases en y intégrant des informations plus précises. Premièrement, nous avons ajouté une information séquentielle qui manquait dans le premier système. Deuxièmement, nous avons aussi ajouté des caractéristiques spatiales locales pour mieux décrire les images des vidéos. Une

extraction de caractéristiques temporelles fut aussi testée dans cette partie. Dans un souci de clarté, et pour intégrer ces différentes améliorations, nous présentons en premier lieu le nouveau système de reconnaissance.

IV.1 Méthodes

Système de reconnaissance

Le processus complet de reconnaissance est présenté sur la **Figure 5**. La première étape de classification supervisée est utile pour extraire des attributs visuels spécifiques à chaque chirurgie. Pour cela plusieurs classifieurs associés à différents type de caractéristiques images (couleur, forme, texture) sont proposés. Une fois que ces attributs visuels ont été détectés, une signature sémantique pour chaque image est ainsi créée. Cette signature sémantique est composée de valeurs représentant les attributs visuels utilisés. La séquence de ces signatures images (i-e: série temporelle) est ensuite traitée par des algorithmes de classification de séries temporelles dans le but d'en déduire l'enchainement des phases de la chirurgie. Nous avons utilisé pour cela les Chaines de Markov Cachée (CMC) et l'algorithme Dynamic Time Warping (DTW) qui permettent de modéliser l'aspect temporel de la chirurgie. En sortie, nous obtenons des séquences correspondant aux différentes étapes chirurgicales. Les paragraphes suivants présentent en détail toutes les étapes de ce processus de reconnaissance automatique.

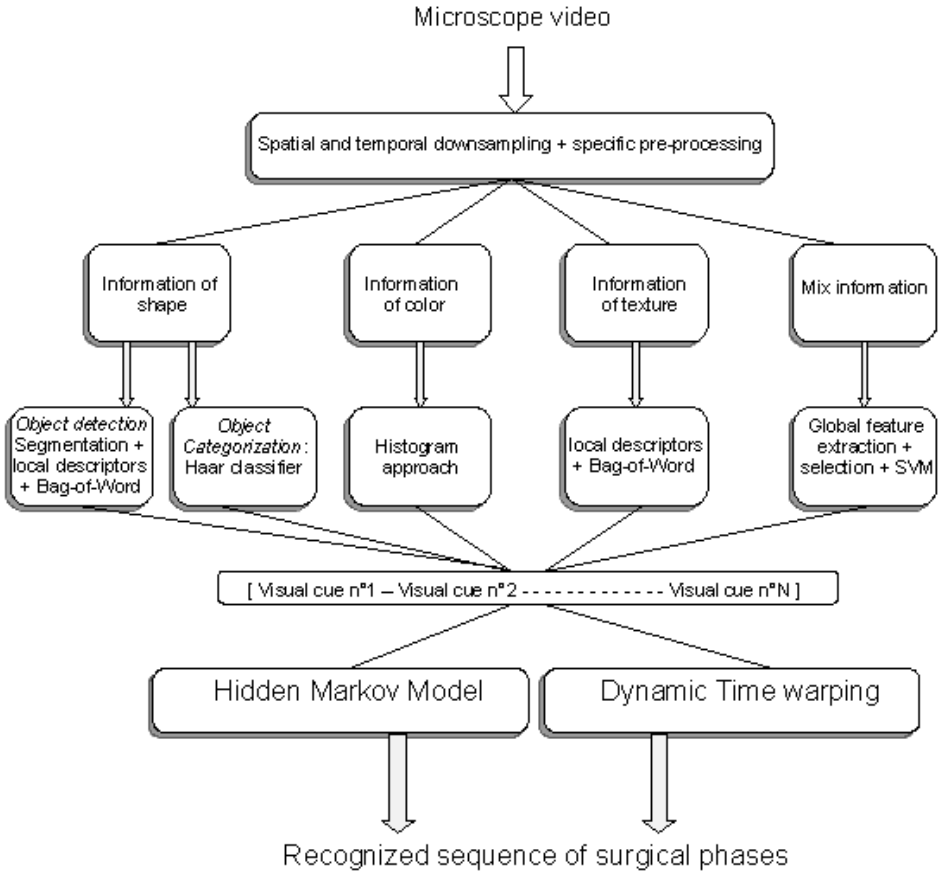


Figure 5 - Nouveau système de reconnaissance de phases.

Pré-traitement

Des étapes de pré-traitement peuvent être nécessaires avant de détecter les attributs visuels. Pour la chirurgie de la cataracte notamment, certains attributs sont identifiables seulement à l'intérieur de la pupille. Dans le but de rendre plus précis la détection de certain attributs, nous avons donc appliqué une segmentation de la pupille basée sur la transformée de Hough (Hough, 1959). Un masque binaire fut d'abord extrait, suivi d'une recherche de cercles par transformée de Hough et d'une normalisation par un diamètre de référence (**Figure 6**).

Figure 6 - Différentes étapes de la segmentation de la pupille.

De même, le nouveau système tend à reconnaitre des informations sur les outils utilisés. Pour cette détection, une première étape de segmentation en Région d'intérêt (ROI) fut effectuée. Pour chaque image, 2 ROIs pouvant correspondre aux 2 outils maximum utilisés par le chirurgien en chirurgie de la cataracte furent extraits grâce à un filtre puis d'une analyse en composantes connexes (**Figure 7**).

Figure 7 - Différentes étapes de la segmentation de la pupille.

Classification des attributs visuels

Cinq classifieurs furent proposés ici pour extraire les attributs visuels de la chirurgie. Chacun de ces classifieurs est lié à un type particulier d'attribut. Les attributs identifiables grâce à leur couleur furent détectés grâce à des histogrammes couleurs. Les attributs identifiables grâces à leur texture furent extraits grâce à une analyse par sac-de-mots couplés à des descripteurs locaux. Pour les objets facilement catégorisables (possédant de forts contours), un classifeur de Haar fut implémenté. Pour les objets difficilement catégorisables, une simple classification binaire (classe outil ou non) fut utilisée. Pour cela, la méthode par sac-de-mots fut appliquée sur les ROIs dans le cas de la chirurgie de la cataracte. Enfin, pour tous les autres attributs visuels qui ne peuvent pas être identifiés à partir uniquement d'une composante couleur, texture ou forme, nous avons utilisé la méthode de classification d'image traditionnelle présentée dans le chapitre précédant.

Définition des attributs visuels et choix du classifieur

L'objectif de cette étape est de définir des attributs visuels qui permettront de discriminer les phases d'une chirurgie particulière. Ces informations, de type binaires, sont considérées comme des attributs propres à chaque chirurgie. En d'autres termes, il a été demandé aux chirurgiens de définir différents attributs visuels pouvant être *a priori* identifiables à travers une analyse d'images. Ces attributs peuvent être, par exemple, la présence/absence d'un objet, le zoom du microscope ou encore une couleur particulière dans la scène chirurgicale. Ces attributs visuels, une fois détectés, sont ensuite extraits en utilisant le classifieur le plus adapté.

Pour la chirurgie de l'hypophyse, 4 attributs visuels ont été défini: le zoom du microscope, la présence des écarteurs nasaux, la présence de la cloison nasale et des compresses. La combinaison de ces quatre attributs binaires permet de discriminer les 6 étapes. Nous avons ensuite choisi le type de classifieur à utiliser pour chaque attribut visuel. Le classifieur de Haar fut utilisé pour détecter les écarteurs nasaux, un histogramme couleur pour la compresse, et la méthode traditionnelle pour les deux autres attributs.

Pour la chirurgie de la cataracte, 5 attributs visuels ont été identifiés : la couleur de l'iris (rouge ou noir), l'aspect globale du cristallin (morcelé ou non), la présence de l'antiseptique, la présence du scalpel et de l'instrument pour implanter la lentille. La couleur de la pupille fut reconnue à partir d'une analyse d'histogramme couleur uniquement sur la pupille après segmentation. Aussi après segmentation de la pupille, l'aspect global du cristallin fut détecté en utilisant une approche par sac-de-mots. La présence de l'antiseptique fut détectée par histogramme couleur sur toute l'image. Pour le scalpel, le classifieur de Haar fut utilisé, ainsi que l'approche par sac-de-mots sur les ROIs pour la détection des instruments. Finalement, l'instrument d'implantation de la lentille fut détecté avec l'approche traditionnelle.

Classification de séries temporelles

Chaine de Markov Cachée (CMC) : Les modèles graphiques probabilistes sont souvent utilisés pour décrire des dépendances entre des données d'observations dans des domaines tels que la biologie. Notamment, les réseaux bayésiens (RB) ont récemment prouvé leur utilité dans ces applications. Les CMC, exemples particuliers de RB, peuvent être utilisées pour modéliser des séries temporelles. Ici, nous utilisons une CMC (Rabiner, 1989) du premier ordre (**Figure 8**) pour modéliser le déroulé de l'intervention. Mathématiquement, une CMC est définie par un 5-uplet (S,O,Π,A,B), où $S=(s_1...(s_N))$ est un jeu fini de N états, $O=(o_1...(o_M))$ est un jeu de M symboles dans un vocabulaire, $\Pi=(\pi(i))$ sont les probabilités d'états initiales, $A=(a(ij))$ sont les probabilités de transitions et $B=(b_i(o(k)))$ les probabilités de sortie. Dans notre approche, les attributs visuels détectés, par la phase de classification supervisée qui précède, sont utilisés comme observations pour la CMC. Ensuite, les différents paramètres du modèle sont déterminés de façon *ad hoc*. Enfin, l'algorithme de Viterbi (Viterbi, 1969) trouve la séquence d'états la plus probable en sortie.

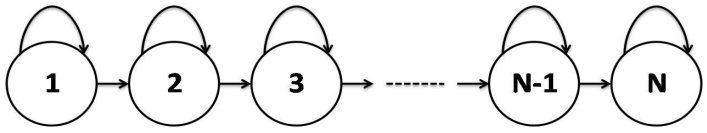


Figure 8 - CMC gauche-droite, où chaque état correspond à une étape de la chirurgie

Algorithme DTW : L'algorithme DTW, décrit parKeogh et Pazzani (1998), est une méthode permettant de classifier une séquence d'images de manière supervisée. Il cherche le chemin optimal entre deux séquences de vecteurs caractéristiques $X = (x_1, x_2, \ldots, x_N)$ de taille N et $Y = (y_1, y_2, \ldots, y_M)$ de taille M. Ces séquences peuvent être des signaux discrets (séries temporelles) ou, de manière plus générale, des séquences de caractéristiques échantillonnées à des points équidistants dans le temps. Pour pouvoir comparer chaque chirurgie, une chirurgie moyenne a été créée à partir de la base d'apprentissage. Ainsi, chaque nouvelle chirurgie est dans un premier temps traitée pour en extraire les caractéristiques visuelles (et créer les signatures images). Ensuite, la séquence des signatures est envoyée à l'algorithme DTW pour être comparée avec la chirurgie moyenne. Une fois la synchronisation effectuée entre les deux séquences, les phases de la chirurgie moyenne sont transposées de manière supervisée à la chirurgie inconnue manipulée. Des contraintes globales (ou fonctions de fenêtrage) peuvent aussi être ajoutées à l'algorithme afin de contraindre les chemins possibles de lien. Ainsi, le chemin ne peut pas sortir de la fenêtre de contraintes. Pour notre modélisation des chirurgies, le choix s'est porté sur le parallélogramme d'Itakura, ce qui permet de conserver notre chemin de lien peu éloigné du chemin diagonal.

Validation

Plusieurs aspects du système de reconnaissance furent validés. Premièrement, la segmentation de la pupille fut comparée à une segmentation manuelle sur le jeu de données entier. Ensuite, l'approche par-sac-de-mots fut optimisée pour les deux classifieurs concernés. Des combinaisons entre différents détecteurs et descripteurs de points-clés furent testées (SIFT, SURF, Harris, STAR, etc.), et le nombre de mots optimal fut déterminé. De la même façon, la reconnaissance des attributs visuels fut analysée. Pour valider ces différents choix de classifieurs, nous avons appliqué la méthode traditionnelle pour détecter chaque attribut des 2 jeux de données et effectuer une comparaison de précision de reconnaissance. Enfin, le taux de reconnaissance global du système fut calculé. Toutes ces expérimentations furent effectuées grâces à des validations croisées.

IV.2 Résultats

Pour la détection de la pupille, une précision de 95 +/- 6% fut trouvée. Pour l'optimisation de l'approche par sac-de-mots pour la détection des instruments, la combinaison points SURF + descripteurs SURF donna les meilleurs résultats pour un nombre de mots égal à 12. De la même façon, l'optimisation de l'approche par sac-de-mots pour l'aspect globale du cristallin, la combinaison points SIFT + descripteurs SURF donna les meilleurs résultats pour un nombre de mots égal à 12. Ces différents paramètres furent gardés pour la suite de l'étude.

La validation des reconnaissances des attributs visuels (**Tableau 2**) montre de très bon résultats, aux alentours de 90% pour les attributs visuels de la chirurgie de l'hypophyse, et aux alentours de 95% pour ceux de la chirurgie de la cataracte. Dans l'ensemble, les classifieurs spécifiques donnent de meilleurs résultats que le classifieur traditionnel, justifiant leur utilisation dans le cadre de cette étude.

Tableau 2 - Précision des reconnaissances des différents attributs visuels, en utilisant les classifieurs spécifiques et le classifieur traditionnel.

	Zoom	Rétracteurs nasaux	Présence cloison nasal	Présence compresse
Classifieur spécifique (%)	88.9 (2.2)	65.2 (8.4)	94.8 (1.3)	87.5 (2.4)
Classifieur traditionnel (%)	X	89.4 (1.1)	X	88.3 (1.6)

	Couleur pupille	Présence antiseptique	Présence scalpel	Présence instrument implantation	Aspect cristallin	Présence instruments
Classifieur spécifique (%)	96.2 (3.6)	96.1 (0.7)	96.7 (3.4)	94.6 (1.1)	87.2 (5.4)	84.1 (8.6)
Classifieur traditionnel (%)	94.1 (4.6)	95.6 (0.4)	88.5 (4.3)	X	54.1 (3.6)	58.7 (6.1)

La validation du système complet de reconnaissance (**Tableau 3**) montre que la reconnaissance des phases est moins bonne pour la chirurgie de l'hypophyse que pour la chirurgie de la cataracte. De même, l'algorithme DTW semble être plus adapté à ce problème de reconnaissance que l'algorithme HMM. Le meilleur taux de reconnaissance est obtenu pour la chirurgie de la cataracte en utilisant l'algorithme DTW (94.4%).

Tableau 3 - Précision de reconnaissance des phases pour les deux jeux de données.

	HMM	DTW
Chirurgie de l'hypophyse	90.2 (6.4)	92.7 (4.2)
Chirurgie de la cataracte	91.4 (5.3)	94.4 (3.1)

Une séquence reconnue par le système, comparée à la séquence réelle est montrée sur la **Figure 9**, pour les deux jeux de données.

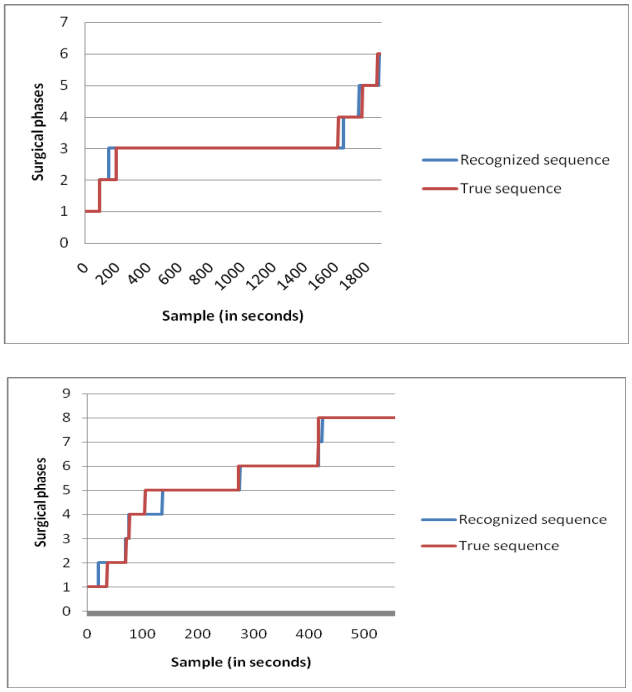


Figure 9 - Séquence reconnue par notre processus de reconnaissance et séquence réelle pour une vidéo :
Haut : jeu de donnée n°1
Bas : jeu de données n°2

IV.3 Discussion

Segmentation de l'iris

Le schéma global de reconnaissance a été développé de manière à être applicable à n'importe quel type de chirurgie. Cependant, comme chaque chirurgie possède ses propres particularités et caractéristiques, il s'avère nécessaire d'effectuer des étapes de prétraitements spécifiques telles que la segmentation de la pupille. Cela va ainsi permettre de pouvoir identifier des caractéristiques visuelles particulières se trouvant à l'intérieur de l'iris. En utilisant notre méthode basée sur une analyse de l'image, la segmentation de la pupille est assez précise. Dans environ 95% des images, la région d'intérêt détectée contient en grande partie la pupille. C'est un résultat suffisant, car l'apport de cette méthode dans le système global permet d'augmenter la détection des phases. Une plus faible précision dans la détection obtenue pour certaines vidéos vient du fait de la présence de rétracteurs dans le champ de vue du microscope optique ce qui restreint la partie visible de la pupille. Il existe d'autres inconvénients similaires qui abaissent la précision de la détection, par exemple lorsque l'iris est trop déformé, lorsque le doigt du chirurgien occulte le champ de vue, lorsque les outils prennent trop de place par rapport à l'iris dans le champ de vue.

Détection des outils chirurgicaux

Notre méthode de détection/reconnaissance des outils chirurgicaux offre des résultats prometteurs: 84,1% de reconnaissance d'un outil chirurgical (sans différenciation de l'outil). C'est un premier résultat assez bon pour cette méthode. Son apport dans la détection des phases est réel mais minime (amélioration de l'ordre de quelques pourcents). Cependant, cette méthode comporte aussi ses inconvénients, principalement au niveau de la détection des composantes connexes (en amont de la reconnaissance). En effet, les masques crées sont automatiques et ne prennent pas en compte les types d'outils chirurgicaux recherchés, ainsi il arrive que seulement des parties des outils soient sélectionnées et non les outils entiers. Cette détection incomplète entraine donc une baisse des résultats pour la reconnaissance. De plus, la classe correspondant à l'arrière-plan (autrement dit, une composante connexe qui n'est pas un outil) a besoin d'un apprentissage plus conséquent du fait de toutes les possibilités qu'elle doit couvrir, ce qui n'est pas optimal car il est difficile de prévoir tous les cas possibles. Effectuer un apprentissage pour chaque classe en validation croisée pourrait permettre d'avoir plus d'images (plus que les 100 actuelles), ce qui devrait ainsi améliorer les résultats de l'algorithme PPV.

Caractéristiques temporelles

En ce qui concerne les méthodes testées pour extraire des caractéristiques temporelles, nous nous sommes focalisés sur le flot optique (Beauchemin and Barron, 1995) les points STIP (Laptev and Lindeberg, 2006). Bien que montrant des résultats satisfaisants, ces méthodes n'ont pour l'instant pas été ajoutées au schéma global. La combinaison entre caractéristiques spatiales et modélisation de série temporelle ne permet pas l'ajout de ce type d'information de mouvement. Dans une future version du ce système, ces méthodes pourraient néanmoins être utiles pour segmenter des zones en mouvement avant la phase de classification des instruments.

Analyse de séries temporelles

Combinée avec des techniques de vision par ordinateur, l'analyse en séries temporelles montre de très bonnes performances et ouvre la voie pour des futurs travaux sur la reconnaissance de tâches de plus haut niveau en chirurgie. L'aspect principal du DTW est qu'il capture l'aspect séquentiel existant entre les phases d'une chirurgie, ce qui est bien adapté à notre étude. La valeur de la fonction de coût entre deux procédures chirurgicales possédant les mêmes enchainements de phases, mais avec des temps passés dans chaque phase différents, sera faible. L'avantage est donc la possibilité de synchroniser de manière précise deux procédures chirurgicales en réduisant les différences de temps passé. L'inconvénient majeur du DTW va survenir dans le cas où l'enchainement des phases n'est pas le même entre deux chirurgies. Dans ces cas, la synchronisation des phases ne se fera pas correctement et des erreurs apparaitront. Une autre limitation de l'algorithme DTW est son impossibilité de fonctionnement en temps réel car la totalité de la procédure chirurgicale est requise pour déterminer le chemin optimal.

De la reconnaissance des phases à la reconnaissance des activités

Comme vu avec ce système, la reconnaissance automatique des phases à partir uniquement des vidéos des microscopes chirurgicaux est tout à fait réalisable. Dans le but de pouvoir utiliser ce genre de systèmes de reconnaissance au sein d'applications cliniques, le challenge maintenant est de descendre d'un niveau de granularité pour détecter les activités de la procédure.

V. Reconnaissance des activités

Une fois la détection des phases effectuée, on s'est intéressé à la détection des activités d'une chirurgie, les activités représentant un niveau de granularité inférieur à celui des phases. Pour cela, il s'avère nécessaire d'utiliser de la connaissance *à priori* en se basant sur la reconnaissance de phases pour restreindre le champ de possibilité des activités lors de l'étape de classification. Dans cette partie nous nous sommes focalisés sur les procédures de chirurgie de la cataracte.

V.1. Méthodes

Comme vu dans le chapitre présentant les jeux de données, une activité est représentée par un triplet du type: <*verbe d'action - outil chirurgical – structure anatomique*>. Ici, le type d'outil utilisé et la zone d'action sont parmi les informations les plus pertinentes pour identifier l'activité correspondante. A l'inverse, le verbe d'action est quasiment impossible à détecter au sein d'une activité, car le mouvement associé est quasiment tout le temps identique (mouvement rectiligne dans la vidéo). Nous nous sommes donc concentrés sur la détection des outils et sur les structures anatomiques. De plus, nous avons intégré à la phase de classification une décomposition hiérarchique de la chirurgie.

Reconnaissance des outils chirurgicaux

Détecter et reconnaitre des outils chirurgicaux peut être relativement complexe du fait de leurs formes similaires, des changements d'illuminations, d'orientation, etc. Ici, nous avons proposé d'étendre l'approche par sac-de-mots, présentée précédemment, à 7 classes, correspondant à 6 types d'outils chirurgicaux et une classe d'arrière-plan. Deux composantes connexes par image furent utilisées, correspondant aux possibles outils chirurgicaux.

Reconnaissance des structures anatomiques

L'autre information utile à extraire est la structure anatomique. Comme l'information de profondeur n'est pas disponible dans les vidéos, 3 zones furent identifiées dans l'image : la zone de la pupille au centre de l'image, la zone de l'iris autour de la pupille, et le reste de l'image. Pour la segmentation, nous nous sommes basés sur la segmentation de la pupille présentée précédemment. Un cercle d'une circonférence de référence correspondant à l'iris fut calculé de la même façon que pour la pupille et ajouté à l'image (**Figure 10**). En utilisant ces informations de zones, et en connaissant la position exacte des différentes ROIs correspondant aux outils chirurgicaux, des pourcentages d'apparitions des outils dans chaque zone peuvent être calculés.

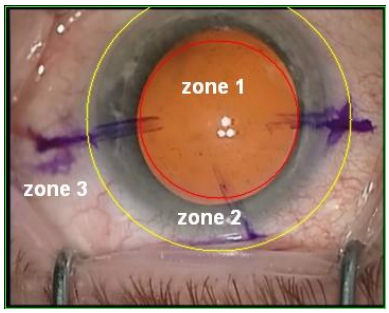


Figure 10 - Illustration des 3 zones: zone 1: pupille, zone 2: iris, zone 3: reste de l'image.

Classification

Cette étape va se faire en injectant de la connaissance en mettant en place une décomposition hiérarchique (que l'on peut assimiler à une ontologie légère sans relation directe entre éléments). Grâce à cette décomposition hiérarchique définissant le lien entre activités et phases, la connaissance préalable des phases permet de restreindre les possibilités et de lancer des classifieurs de type supervisés (ici PPV) sur un nombre restreint de classes (couples d'activités). L'algorithme DTW fut utilisé pour la détection préalable des phases. Les résultats de la classification DTW n'étant pas parfaits, il existe un décalage entre la phase reconnue et la phase véritable, que l'on prend en compte en ajoutant lors de l'étape de classification les paires d'activités de la phase précédente ainsi que celles de la phase suivante.

V.2 Résultats

Tableau 4 - Taux de reconnaissance des activités, spécificité and sensibilité.

	Moyenne (std)	Spécificité	Sensibilité
Taux de reconnaissance (%)	64.5 (6.8)	54.9	76.3

Le taux de reconnaissance des activités chirurgicales fut de l'ordre de 64% (**Tableau 4**). De même, des études sur les taux de reconnaissance par activités furent calculés et permirent de s'apercevoir que certaines paires d'activités étaient facilement reconnaissables (précision > 95%) quand d'autres étaient quasiment impossibles à détecter (précision < 10%).

V.3 Discussion

Le choix d'effectuer la reconnaissance des activités en utilisant d'une part des caractéristiques visuelles et d'autre part des informations provenant de l'alignement DTW (utilisé pour la reconnaissance des phases) semble être opportun. En effet, les résultats préliminaires obtenus de cette manière permettent à notre système d'obtenir des taux de bonne reconnaissance des activités de l'ordre de 65%. Même si les résultats ne sont pas parfaits, cela permet d'avoir une bonne idée de l'enchainement des activités effectuées par le chirurgien au cours de la procédure et ce de manière automatique.

L'utilisation d'un classifieur basique de type PPV avec des signatures images semble être une bonne solution pour effectuer la reconnaissance des activités car les résultats sont encourageants. Cependant, les valeurs prises en compte pour créer ces signatures images ne sont pas forcément optimales. Utiliser des informations provenant des composantes connexes et donc par extension relatives aux outils chirurgicaux manipulés est logique car une activité est directement liée à l'utilisation d'un outil. Toutefois, les imperfections liées à notre méthode de reconnaissance des outils se répercutent ici et entrainent une baisse des résultats. De plus, créer une seule signature image à partir des informations provenant des différentes composantes connexes et de la première passe de reconnaissance des phases est une tâche complexe. C'est en partie ce qui explique la limite de notre système concernant la reconnaissance automatique des activités. Toutes ces informations sont utiles et nécessaires pour l'identification des activités, il faut cependant réfléchir à d'autres manières de les mettre ensemble dans une seule signature image.

VI. Discussion et conclusion

La première partie de la thèse s'est focalisée sur la reconnaissance automatique des phases chirurgicales. L'idée sous-jacente fut de combiner des techniques de vision par ordinateur robustes permettant d'extraire des attributs visuels, à une analyse de séries temporelles pour prendre en compte l'aspect séquentiel des phases. Premièrement, des attributs visuels pertinents qui permettent de discriminer les différentes phases de la chirurgie furent manuellement définis. Cinq classifieurs furent implémentés pour reconnaitre ces attributs dans l'image, où chacun de ces cinq classifieurs fut relié à un type de caractéristiques à extraire. Les attributs reconnaissables à travers leurs couleurs furent extraits avec des histogrammes couleur. Pour les attributs reconnaissables à travers leur forme, deux types de classifieurs furent implémentés. Le premier fut un classifieur de Haar pour catégoriser des objets de fort contour. Le deuxième fut une approche par sac-de-mots pour détecter des objets, sans pour autant arriver à les catégoriser. Les attributs reconnaissables à travers leur texture furent aussi appréhendés par une approche par sac-de-mot. Enfin tous les autres attributs ne rentrant pas clairement dans un de ces types de caractéristiques, ou encore en associant plusieurs, furent reconnus grâce à un classifieur standard mêlant extraction de caractéristiques bas-niveau, sélection de caractéristique et classification supervisée. Cette première étape de traitement d'image pur permet de caractériser chaque frame de la vidéo de façon statique, et après concaténation des signatures images de créer des séries temporelles qui peuvent être ensuite présentées en entrée de système d'analyse de séries temporelles. Nous avons implémenté deux de ces méthodes, les chaînes de Markov Cachées et le *Dynamic time Warping*.

La seconde partie de la thèse s'est focalisée sur la reconnaissance automatique des activités chirurgicales (tâches bas-niveau). Ce niveau de granularité est formalisé par des triplets < *action* – *outil chirurgical* – *structure anatomique* >. Des informations plus précises sur les outils chirurgicaux

ainsi que sur les zones d'apparitions de ces outils furent mixées avec les attributs visuels précédemment extraits au sein de signatures images plus détaillées. Ensuite, en se basant sur l'hypothèse que la plupart des activités apparaissent seulement dans une ou deux phases, une décomposition hiérarchique de la procédure fut créée pour faire le lien entre phase et activité. En utilisant cette décomposition hiérarchique, les résultats de la classification des phases et les nouvelles signatures images, la classification des activités devient possible. Des précisions globales de reconnaissance des phases et des activités ont été calculées, en se basant sur deux jeux de données : un jeu de données de vidéos de neurochirurgie, et un jeu de données de vidéos de chirurgie ophtalmologique. Nous avons obtenus des résultats de l'ordre de 95% pour la reconnaissance des phases de la chirurgie et de l'ordre de 65% pour la reconnaissance des activités. Ces résultats sont très bons pour la reconnaissance des phases et encourageant en ce qui concerne la reconnaissance des activités.

Un des axes principaux de cette thèse fut l'utilisation de vidéos des microscopes chirurgicaux en entrée des systèmes de reconnaissance. Comme expliqué par Bouarfa et al. (2010), les informations extraites des salles d'opération doivent être discriminantes, facilement contrôlables, invariantes selon le chirurgien qui opère et ne pas demander beaucoup de ressources. Les données vidéos issues des microscopes réunissent toutes ces contraintes. Elles sont tout d'abord assez discriminantes pour pouvoir obtenir des taux de classification de l'ordre de 90% pour la détection des attributs images. La performance de cette classification est largement liée à la puissance de discrimination de la base de données images. Cette contrainte soulève le problème de l'adaptabilité du système pour les différents hôpitaux. Au sein d'un même service, les couleurs des instruments ou les particularités de la scène chirurgicale sont les mêmes, mais pour d'autres services hospitaliers, cela peut ne pas être le cas. Par exemple, la couleur des tissus utilisés à Rennes peut ne pas être la même que celle d'autres services. La solution serait alors d'entrainer une base de données pour chaque service, qui prendrait en compte l'environnement local. Un autre problème pourrait se poser avec la variabilité entre les chirurgiens. Mais en supposant que l'environnement est le même pour chaque chirurgien au sein d'un même service, cette variabilité est réduite au temps de chirurgie, qui ne biaise pas du tout notre modèle. Ensuite, les données sont toujours disponibles selon le même format, ce qui facilite l'utilisation qui va en être faite derrière. Enfin, l'atout principal de ce projet est l'utilisation du microscope. Cet appareil est non seulement déjà installé dans les salles d'opération, mais en plus l'enregistrement n'a pas à être contrôlé par une personne pendant l'intervention. Ce type de données est donc une bonne solution pour la création de nouveaux systèmes de chirurgie guidée par l'image incluant une connaissance explicite et formalisée de l'activité chirurgicale en cours.

Ces systèmes de reconnaissance de tâches chirurgicales, que ce soit au niveau des phases ou au niveau des activités, apparaissent comme une progression non négligeable vers la construction de systèmes intelligents (autrement dit sensibles au contexte) pour la chirurgie. Dans leurs versions actuelles, les systèmes peuvent être utilisés en postopératoire afin d'indexer les vidéos en fin de chirurgie et de créer des rapports chirurgicaux pré-remplis. Dans le cadre de l'enseignement, avoir à disposition une base de données de vidéos chirurgicales indexées peut être aussi utile et une navigation entre les différentes phases et activités des chirurgies pourrait être effectuée. Une des perspectives principales de cette thèse est l'utilisation de systèmes équivalents dans les salles d'opération en temps-réel. Pour le moment, certains algorithmes (DTW par exemple) ne fonctionnent que lorsque la vidéo est entièrement terminée, ce qui limite les champs d'application du système. Une des applications temps-réel qui pourrait être amenée à voir le jour est l'assistance intra-opératoire, par exemple en

permettant en temps réel de savoir quelles informations ont besoin d'être montrées au chirurgien pour la tache effectuée. Cela pourrait aussi permettre une meilleure anticipation de possibles évènements néfastes permettant d'optimiser la chirurgie et de réduire les dangers pour le patient. Les systèmes de reconnaissances basés sur les vidéos des microscopes, que ce soit pour la détection des tâches de haut ou bas-niveau, offrent donc de réelles perspectives d'avenir dans le domaine de la CAO.

Références

- ✓ Ahmadi A, Sielhorst T, Stauder R, Horn M, Feussner H, Navab N. Recovery of surgical workflow without explicit models. Proc MICCAI, Berlin: Springer. 2007; 420-8.
- ✓ Ahmed N, Natarajan T, Rao KR. Discrete Cosine Transform. IEEE Trans Comp. 1974. 90-3.
- ✓ Bhatia B, Oates T, Xiao Y, Hu P. Real-time identification of operating room state from video. AAAI.2007; 1761-6.
- ✓ Beauchemin SS, Barron JL. The computation of optical flow. ACM New York, USA. 1995.
- ✓ Darzi A, Mackay S. Skills assessment of surgeons. Surgery. 2002; 131(2): 121-4.
- ✓ James A, Vieira D, Lo BPL, Darzi A, Yang GZ. Eye-gaze driven surgical workflow segmentation. Proc MICCAI. 2007; 110-7.
- ✓ Duda, R.O. and Hart, P.E. Pattern classification and scene analysis. Guyon, John Wiley & Sons. 1973.
- ✓ Hamming, R.W. Coding and Information Theory. Prentice-Hall Inc. 1980.
- ✓ Haralick, RM., Shanmugam, K., Dinstein, I. Textural features for image classification. IEEE Trans. on Systems, Man, and Cybernetics. 1973; 3(6): 61-2.
- ✓ Hough VC. Machine Analysis of Bubble Chamber Pictures. Proc Int Conf High Energy Accelerators and Instrumentation. 1959.
- ✓ Hu MK. Visual pattern recognition by moment invariants. IRE Trans on Information Theory. 1962; 8(2): 179-187.
- ✓ Jannin P Morandi X. Surgical models for computer-assisted neurosurgery. Neuroimage. 2007; 37(3): 783-91.
- ✓ Keogh EJ, Pazzani MJ. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. Prediction of the future: AI approaches to time-series problems. 1998; 44-51.
- ✓ Klank U, Padoy N, Feussner H, Navab N. Automatic feature generation in endoscopic images. Int J Comput Assist Radiol Surg. 2008; 3(3,4): 331-9.
- ✓ Laptev I, Lindeberg T. Local descriptors for spatio-temporal recognition. Spatial coherence for visual motion analysis. 2006. Springer
- ✓ Lin HC, Shafran I, Yuh D, Hager GD. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Computer Aided Surgery. 2006; 11(5): 220-30.
- ✓ Lo B, Darzi A, Yang G. Episode Classification for the Analysis of Tissue-Instrument Interaction with Multiple Visual Cues. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2003.
- ✓ Nara A, Izumi K, Iseki H, Suzuki T, Nambu K, Sakurai Y. Surgical workflow monitoring based on trajectory data mining. New frontiers in Artificial Intelligence. 2011; 6797: 283-91.
- ✓ Neumuth T, Trantakis C, Eckhardt F, Dengl M, Meixensberger J, Burgert O. Supporting the analysis of inter-vention courses with surgical process models on the example of fourteen microsurgical lumbar discectomies. Int J Comput Assisted Radiol Surg. 2007; 2(1): 436-8.
- ✓ Padoy N, Horn M, Feussner H, Berger M, Navab N. Recovery of surgical workflow: a model-based approach. Int J Comput Assisted Radiol Surg. 2007; 2(1): 481-2.

- ✓ Rabiner LR. A tutorial on Hidden Markov Models and selected applications in speech recognition, Proc of IEEE. 1989; 77(2).
- ✓ Speidel S, Sudra G, Senemaud J, Drentschew M, Müller-stich BP, Gun C, Dillmann R. Situation modelling and situation recognition for a context-aware augmented reality system. Progression in biomedical optics and imaging. 2008; 9(1): 35.
- ✓ Smeulders AW, Worrin M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Trans on pattern analysis and machine learning intelligence. 2000; 22(12): 1349-80.
- ✓ Xiao Y, Hu P, Hu H, Ho D, Dexter F, Mackenzie CF, Seagull FJ. An algorithm for processing vital sign monitoring data to remotely identify operating room occupancy in real-time. Anesth Analg, 2005; 101(3): 823-2.
- ✓ Viterbi A. Errors bounds for convolutional codes. IEEE TIT. 1967; 13(2): 260-9.

ANNEXE 2 (Modèle dernière page de thèse)

VU :

Le Directeur de Thèse
(Nom et Prénom)

VU :

Le Responsable de l'École Doctorale

VU pour autorisation de soutenance

Rennes, le

Le Président de l'Université de Rennes 1

Guy CATHELINEAU